ACADEMIE DE MONTPELLIER

UNIVERSITE MONTPELLIER II
(Sciences et Techniques du Languedoc)

D.E.S.S.
METHODES STATISTIQUES DES INDUSTRIES AGRONOMIQUES, AGROALIMENTAIRES ET PHARMACEUTIQUES

MEMOIRE sur le stage

Utilisation de modèles de régressions logistique et log-binomiale dans l'étude du retard de croissance en taille chez le jeune enfant sénégalais en milieu urbain

- effectué du 1er Mars 1999 au 31 Août 1999

- au

Laboratoire de Nutrition Tropicale – IRD
(Institut de Recherche pour le Développement, anciennement Orstom)
DAKAR-MONTPELLIER
Sous la direction d'Agnès Gartner et Pierre Traissac

- par M. **Thierry HOARAU**

- soutenu le 14 septembre 1999

- devant la commission d'examen

A. DELCAMP, A. GANNOUN, R. SABATIER, J. SARACCO

REMERCIEMENTS

Je tiens tout d'abord à remercier messieurs les Professeurs Alain Delcamp et Ali Gannoun, co- responsables du DESS MSIAAP, pour m'avoir permis de suivre cette formation.

Je remercie également les Docteurs Francis Delpeuch, directeur du Laboratoire de Nutrition Tropicale IRD et Bernard Maire, responsable du programme "Dynamiques nutritionnelles en milieu urbain", pour m'avoir accueilli au sein de cette unité.

Un grand merci à Pierre Traissac et Agnès Gartner pour leur patience, leur bonne humeur et leur encadrement en tant que maîtres de stage. J'ai beaucoup appris avec vous et espère en apprendre davantage dans le cadre de mon futur poste en tant que CSN à Dakar.

J'adresse de chaleureux remerciements à l'équipe de nutrition de Dakar grâce à qui mon séjour a été des plus agréables. Merci à vous Denis, Marie-Claire, Virginie, Yves, Vincent, Toffène, Pap, Amy, Nicolas, Gnagna, Ouli et Omar.

Je remercie aussi toute l'équipe de nutrition de Montpellier et en particulier Catherine Marchand, Marie-Catherine Vieu et Antoine Foltzer pour leur sympathie.

Enfin je tiens à adresser une pensée toute particulière à ma famille qui m'a toujours soutenu malgré la distance (île de la Réunion).

Cadre du travail et objectifs généraux

Cette étude a été menée dans le cadre du programme "Dynamiques nutritionnelles en milieu urbain" du Laboratoire de Nutrition Tropicale de l'IRD (Institut de Recherche pour le Développement, anciennement Orstom), centre collaborateur de l'OMS pour la nutrition dirigé par Francis Delpeuch.

L'IRD, établissement public à caractère scientifique et technologique, conduit des activités de recherche finalisée vers le développement dans une trentaine de pays des régions chaudes de la planète : en Afrique, en Amérique latine, dans le Pacifique et, plus récemment, en Asie tropicale, ainsi que dans les Dom-Tom. Le Laboratoire de Nutrition Tropicale oriente plus précisément ses actions vers la recherche de solutions pour l'amélioration de la situation nutritionnelle des populations dans un contexte de pauvreté et de ressources limitées.

Un des aspects de la problématique porte sur les variations de l'état nutritionnel des populations dans un contexte de crise économique et notamment dans les quartiers pauvres des villes du Sénégal. L'objectif est de mieux cerner les déterminants de la malnutrition, en particulier mieux comprendre les liens entre pauvreté et état nutritionnel selon les situations sociales, et d'améliorer les interventions.

Dans ce cadre une enquête a été menée à Pikine, banlieue de Dakar, en 1996 afin d'évaluer la situation nutritionnelle des enfants de moins de 5 ans et de leur mère et de dégager un ensemble de facteurs explicatifs liés à cet état nutritionnel. Parmi les problèmes nutritionnels évalués, il a été décidé pour ce stage d'étudier le retard de croissance en taille chez les enfants de moins de 5 ans.

Les objectifs de mon travail ont consisté, d'une part, à étudier, en continuité de l'analyse descriptive uni- et bivariée déjà menée par les chercheurs (Unité de Nutrition de l'Orstom, 1997), les facteurs de risque du retard de taille par l'utilisation de méthodologies d'analyse multivariée. Selon leur nature, de tels facteurs peuvent être utiles au dépistage de groupes à risque, au suivi des modifications de situations nutritionnelles ou à l'évaluation des résultats d'intervention. Il sera intéressant, d'autre part, de laisser au terme de ce travail des programmes et des macros utilisables éventuellement pour des études similaires.

La première partie du stage s'est déroulée de mars à juin 1999 au Laboratoire de Nutrition du Centre IRD de Dakar au Sénégal, sous la responsabilité d'Agnès Gartner, soit dans l'équipe qui a mis en place et effectué le recueil de données. Elle a porté sur la prise en main de l'outil SAS, la gestion des données et les premières analyses uni- bi- et trivariées. Le fait de travailler à Dakar a permis de retourner dans les dossiers d'enquête pour d'ultimes mises au point sur les données, mais aussi de connaître le cadre du travail sur le terrain, le contexte sénégalais et la réalité de Pikine, la zone urbaine étudiée.

La deuxième partie du stage a eu lieu en juillet et août 1999 au Laboratoire de Nutrition Tropicale du Centre IRD de Montpellier sous la responsabilité de Pierre Traissac, afin de perfectionner les analyses multivariées et finaliser la rédaction du mémoire

SOMMAIRE

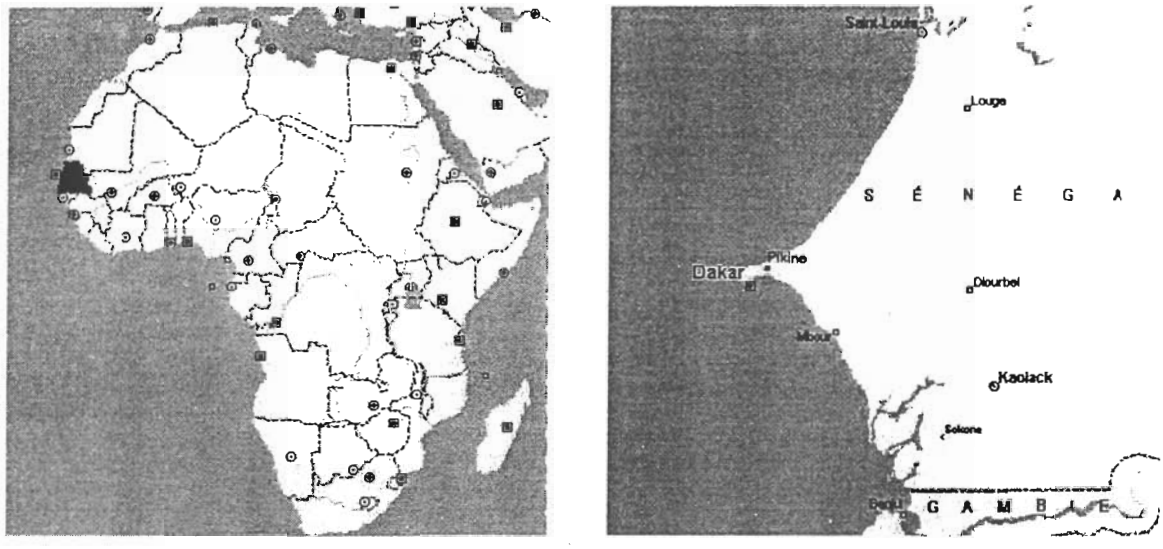
Liste des Formules

Liste des Equations de modèles

Liste des Annexes

Liste des Tableaux

INTRODUCTION



Cartes 1 et 2 : Localisation du Sénégal et de Pikine (source Atlas mondial Microsoft ® Encarta ® Edition 1998)

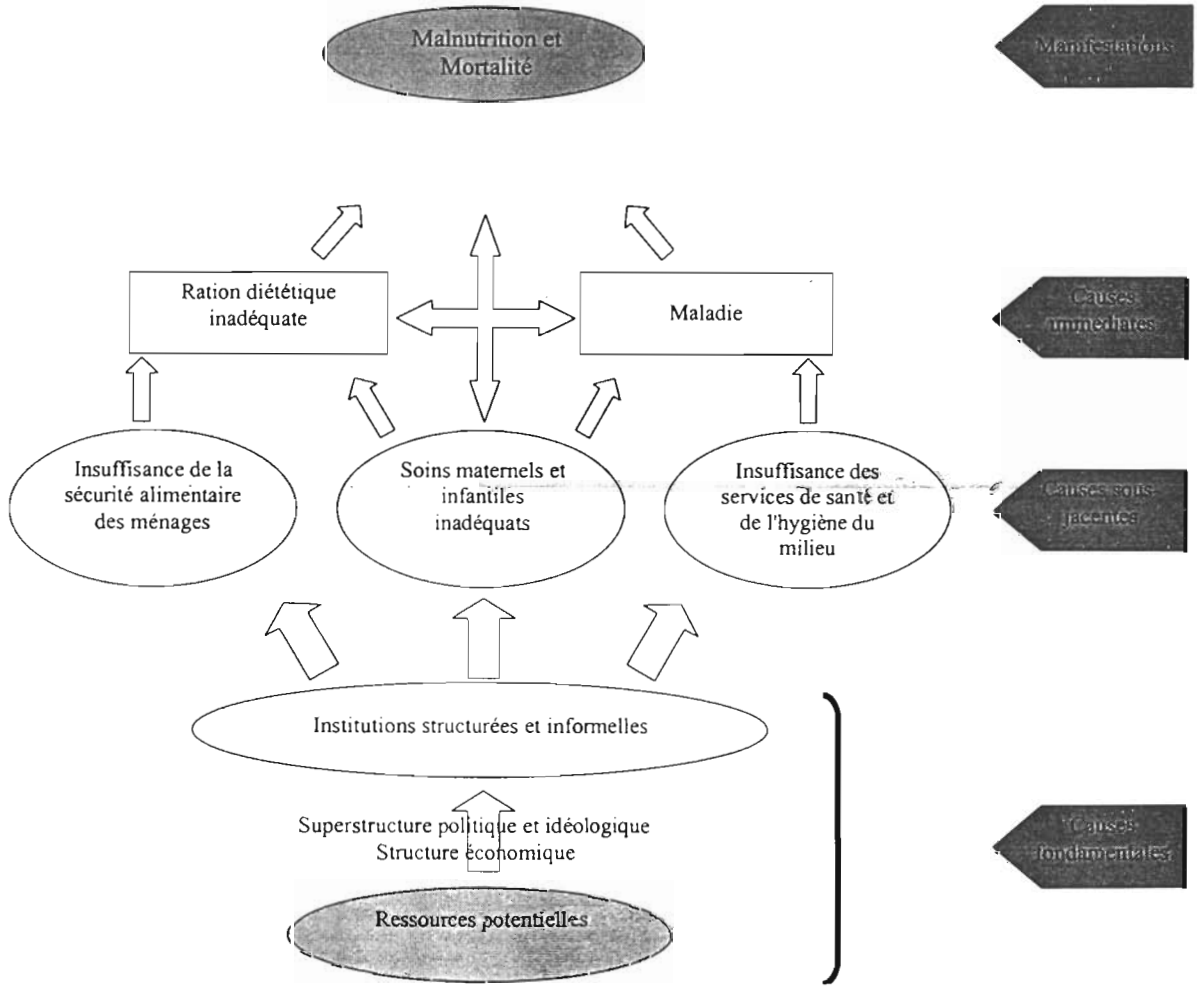


Figure 1 : Diagramme causal de la malnutrition chez le jeune enfant (source FAO, 1997)

1) Situation du Sénégal[1]

Limité au Nord par la Mauritanie, au Sud par la Guinée Bissau et la Guinée, à l'Ouest par l'océan Atlantique et à l'Est par le Mali, le Sénégal couvre une superficie totale de 196 720 km² (**Carte 1**). Le relief est peu marqué sur la majeure partie du pays qui compte quatre grandes zones climatiques : la zone subsahélienne avec une pluviométrie inférieure à 350 mm, la zone sahélienne[2] avec une pluviométrie entre 350 et 500 mm, la zone soudano-sahélienne entre 500 et 800 mm et la zone soudanienne à sub-guinéenne avec des précipitations entre 800 et 1300 mm. La population était estimée à quelques 8 311 000 habitants en 1995, dont 56% de ruraux (contre 78% en 1992). Entre 1990 et 1995, le taux annuel d'accroissement démographique a été de 2.5% en moyenne (0.4% pour la France), mais la population urbaine a augmenté plus vite avec un taux de 4.3%, alors que l'accroissement de la population rurale a été relativement modéré avec un taux de 1.9%. L'économie sénégalaise est fortement dépendante de l'agriculture et de l'exportation de l'arachide, dont la production est tributaire des aléas climatiques et des variations des cours mondiaux. En 1992 le PIB par habitant était estimé à 812 $ US (23 149 $ US pour la France). L'indice de développement humain[3] en 1999 classait le Sénégal 153/174 (7/174 pour la France). De plus les indicateurs de l'état nutritionnel (23% des enfants de moins de 5 ans ont un retard de taille et 7% sont atteints de maigreur)[4] montrent qu'il existe un problème de malnutrition, notamment chronique, à travers le retard de croissance en taille des jeunes enfants.

2) Le retard de taille chez le jeune enfant

La taille de l'enfant reflète la croissance linéaire passée et donc les conditions de croissance depuis la naissance. Le retard de taille est défini par une taille significativement plus basse que la médiane des enfants de même âge de la population de référence NCHS[5] (OMS, 1995). Il indique que le sujet n'a pas réussi à atteindre son potentiel de croissance linéaire en raison de conditions de santé et de nutrition défavorables. La plupart des facteurs économiques, culturels et sociaux qui peuvent expliquer l'apparition du retard de taille sont aussi ceux qui peuvent produire des conséquences néfastes pour l'enfant (altération du développement psychomoteur, survenue des infections, mortalité). Le retard de taille peut être un indicateur de déficits cumulés de la croissance et/ou d'une altération passée de la croissance.

Dans beaucoup de pays en développement, la prévalence du retard de croissance commence à augmenter vers l'âge de 3 mois; le processus de retard de croissance se ralentit vers l'âge de 3 ans, après quoi la taille moyenne évolue parallèlement à la courbe de référence.

Le retard de croissance pendant l'enfance peut aboutir à une réduction significative de la taille de l'adulte. Une des principales conséquences est une diminution de la capacité de travail. D'autre part la taille de la mère est associée à certaines conséquences en matière de reproduction. Les femmes de petite taille présentent un risque augmenté de complications obstétricales. Il existe aussi une forte association entre la taille de la mère et le poids de naissance; il s'ensuit un effet d'une génération à l'autre car les enfants de faible poids de naissance risquent de présenter ultérieurement des déficits anthropométriques.

---

[1] Données extraites de l'Atlas Mondial Microsoft ® Encarta ® Edition 1998

[2] Zone où se situe Dakar (**Carte 2**)

[3] Indice composite incluant l'espérance de vie, le taux d'alphabétisation, le taux de scolarisation et le revenu per capita Source http //www undp.org/hdro/HDI.html

[4] Données Unicef (Fonds des Nations Unies pour l'enfance) http unicef org/statis/country_1page156 html

[5] National Center for Health Statistics

Tableau 1 : Prévalences du retard de taille selon l'âge de l'enfant

Age (mois)	n		Prévalences (%)
0-11	883	20%	6.1
12-23	938	21%	15.6
24-35	933	21%	13.2
36-47	879	20%	16.9
48-59	844	19%	12.9
Total	4477	100%	12.9

3) Facteurs de risque du retard de taille (Figure 1)

Parmi les principaux déterminants conduisant à une malnutrition, certains agissent directement et d'autres indirectement. Les facteurs directs sont l'insuffisance de l'apport alimentaire et la maladie. Certains facteurs indirects sont de nature socio-économique. Par exemple, la pauvreté peut se traduire par un faible niveau d'instruction des parents, de mauvaises conditions d'approvisionnement en eau et assainissement, un manque de moyens pour acheter de la nourriture, un manque d'hygiène conduisant à une contamination des aliments, et des soins de santé insuffisants; tous ces facteurs contribuent à accroître le risque de maladie et à réduire l'absorption d'aliments énergétiques et de nutriments. Les facteurs culturels, influencés par l'environnement social et économique, jouent également un rôle important dans l'étiologie des troubles de la croissance. A titre d'exemple on peut citer la façon d'élever les enfants, y compris les tabous alimentaires. Mis à part la situation économique générale d'une population, les causes d'une forte prévalence de retard de taille sont difficiles à cerner. De plus ces causes sont spécifiques à chaque contexte. Les études épidémiologiques sont utiles pour identifier ces facteurs de risque indirects au niveau de la population. L'existence de liens entre la taille et le niveau socio-économique a été démontrée. C'est pourquoi dans notre étude nous avons retenu un nombre important de variables décrivant l'environnement et les caractéristiques socio-économiques du ménage où vit l'enfant.

4) Evaluation du risque

Le retard de taille pouvant être expliqué par de multiples causes, il s'agit de mettre en place une analyse multifactorielle afin de compléter l'analyse descriptive uni- et bivariée déjà menée sur un échantillon représentatif de la population.

Les études d'observation portant sur les déterminants des déficits anthropométriques présentent quelques difficultés méthodologiques. La question des facteurs de confusion est particulièrement importante dans la mesure où les déterminants de l'état nutritionnel ne sont pas indépendants les uns des autres. L'analyse rigoureuse et la prise en compte des facteurs de confusion et des facteurs modificateurs d'effet tout en tenant compte de la vraisemblance biologique permettent d'obtenir des informations sur les déterminants des malnutritions.

Le risque de maladie induit par un ou plusieurs facteurs peut être quantifié par des mesures dites d'association telles que l'odds-ratio ou le risque relatif. La régression logistique est le modèle "traditionnellement" utilisé pour répondre à cet objectif dans le cas d'une maladie décrite par une variable à deux classes. Toutefois lorsque la maladie étudiée n'est pas rare dans la population, ce modèle a tendance à surévaluer l'importance des facteurs de risque si l'on interprète directement les odds-ratios comme des risques relatifs. Une question à laquelle on tentera de répondre est de savoir si une prévalence moyenne de 13% pour le retard de taille chez les enfants de moins de 5 ans a des effets sur les estimations faites par le modèle logistique (**Tableau 1**). C'est pourquoi une comparaison des résultats sera faite avec ceux d'un autre modèle peut-être plus adapté : la régression log-binomiale.

METHODES

A. LES DONNEES UTILISEES

- Origine des données

L'enquête s'est déroulée en mai et juin 1996, soit avant l'hivernage (saison des pluies). Il s'agit d'une enquête transversale représentative sur l'ensemble de l'agglomération de Pikine, menée auprès des ménages par passage à domicile. Les ménages inclus ont au moins un enfant de moins de 5 ans. La seule cause de non-inclusion est le refus du ménage de participer à l'enquête au moment de la constitution des grappes. Il n'y a aucune cause de non-inclusion spécifique pour l'enfant ou sa mère. La zone a été divisée en 7 strates géographiques afin de prendre en compte la diversité des situations urbaines. A partir de points de départ tirés au sort sur des cartes au 1/2000, 170 grappes ont été constituées sur le terrain par progression par proximité dans les concessions et dans des directions tirées au sort. Les données socio-économiques, familiales, sanitaires et alimentaires, ainsi que les mesures anthropométriques (poids, taille, âge) des enfants de moins de 5 ans et de leur mère ont été recueillies sur un échantillon représentatif final de 2268 ménages, 3150 mères et 4591 enfants. La méthodologie complète et la description du recueil des données sont précisées par ailleurs (Unité de Nutrition de l'Orstom, 1997). Les dossiers de l'enquête ont fait l'objet, par le personnel de l'unité de Nutrition à Dakar, d'une double saisie suivie d'une validation des données mises en œuvre à l'aide du logiciel Epi-Info version 6.01 (Dean et al., 1994). Après exportation les données ont été récupérées pour être traitées avec le logiciel SAS version 6.11 pour ordinateur PC compatible (SAS Institute Inc., 1989 a).

A l'issue de la saisie des données on a dénombré :
- 5 enfants avec des données biologiquement improbables
- 8 nouveau-nés ayant des mesures de poids trop faibles pour pouvoir calculer des indices nutritionnels. L'échantillon est alors de 4578 enfants de moins de 5 ans.

- La variable à expliquer : le retard de taille

Au cours de l'enquête, la taille a été mesurée au mm près, couchée pour les enfants de moins de 24 mois et debout pour les enfants de 24 à 59 mois, à l'aide de toises en bois. La détermination de l'âge précis de l'enfant a été réalisée par vérification de la date de naissance sur un document civil ou un carnet de santé dans plus de 85% des cas. Lorsque celle-ci n'a pas pu être vérifiée, elle a été estimée en utilisant un calendrier des événements locaux. A partir de ces deux paramètres, on compare la taille pour l'âge aux tailles de référence NCHS pour le même âge (OMS, 1995) par la méthode des z-scores. L'indice déduit est l'écart entre une valeur individuelle et la médiane de la population de référence, divisé par l'écart type de la population de référence :

$$Indice_{taille-âge} = \frac{(\text{valeur observée}) - (\text{médiane de référence})}{\text{écart - type de la population de référence}}$$ **Formule 1 : Indice taille-âge**

Cet indice a été calculé selon le sexe de l'enfant à l'aide du logiciel Epinut (Dean et al., 1994). On définit un retard de taille lorsque cet indicateur numérique a une valeur inférieure à -2 (taille telle que dans la population de référence seulement 2.3% des enfants de même âge ont une taille inférieure). Le codage de cet indice continu en indice à deux classes permet,

Tableau 2 : Liste des variables proposées pour l'étude du retard de taille

Caractéristiques du ménage (8)

- Taille du ménage[1]
- Nombre d'enfants de moins de 5 ans[1]
- Densité d'individus par pièce[1]
- Statut d'occupant du ménage
- Source d'eau[2]
- Quantité d'eau disponible
- Sanitaires[2]
- **Indice de niveau économique (12)[1]** ⇐ AFC

- Possession de mouton avant tabaski
- Possession de mouton après tabaski
- Matériau des murs
- Matériau du toit
- Matériau du sol
- Source d'eau
- Electricité
- Sanitaires
- Combustible domestique
- Qualité du logement
- Nombre de pièces
- **Indice de biens (14)**

Somme pondérée des 14 biens ci-dessous

- 1*Radio
- 2*Réchaud
- 5*RadioK7
- 5*Ventilateur
- 10*Réfrigérateur
- 15*Congélateur
- 15*Salon
- 15*Cuisinière
- 20*Téléphone
- 30*Télévision
- 30*Mobylette
- 25*Magnétoscope
- 50*Générateur
- 100*Véhicule

Caractéristiques du chef de ménage (6)

- Age[1]
- Sexe
- Ethnie[2]
- Niveau scolaire[2]
- Secteur d'activité[2]
- Durée de résidence à Pikine[1,3]

Caractéristiques de la mère (11)

- Mère biologique
- Taille[1]
- Age[1]
- Ethnie[2]
- Niveau scolaire[2]
- Nombre d'enfants de moins de 5 ans[1]
- Situation matrimoniale[2]
- Occupation[2]
- Parenté avec le chef de ménage[2]
- Enfant hospitalisé pour malnutrition grave
- Allocations familiales

Caractéristiques de l'enfant (13)

- Sexe
- Age[1]
- Connaissance du poids de naissance
- Rang de l'enfant dans la fratrie[1]
- Parenté avec le chef de ménage
- Existence de tabous alimentaires
- Dernière diarrhée récente[4]
- Dernière diarrhée traitée avec médicaments
- Dernière diarrhée traitée avec pain de singe[5]
- Dernière diarrhée traitée avec la solution sucrée/salée
- Dernière diarrhée traitée avec sachet Unicef[6]
- Existence d'un carnet de vaccination
- Existence d'une fiche de suivi de croissance

[1] Variables quantitatives "discrétisées"
[2] Variables qualitatives avec regroupement de modalités
[3] Localisation voir **Carte 2**
[4] Au cours des deux dernières semaines
[5] Fruit du baobab
[6] Fonds des Nations Unies pour l'Enfance

d'une part, de classer un enfant comme ayant un retard de taille ou non et, d'autre part, de décrire notre échantillon d'enfants par une prévalence de retard de taille (12.9% pour notre échantillon). La variable à expliquer est donc qualitative à deux classes. Selon les recommandations de l'OMS (1983), les enfants ayant un indice <-5 ou >+3 ont été exclus de l'analyse. Onze enfants sont dans ce cas et l'échantillon est alors de 4567 enfants de moins de 5 ans.

- Les variables explicatives proposées : les facteurs de risque potentiels

Dans cette étude il n'y a pas un facteur d'exposition spécifique ou unique et d'autres facteurs de risque mais un ensemble de données décrivant l'environnement socio-économique, sanitaire et familial de l'enfant. Ces variables décrivent le ménage, le chef de ménage, la mère et l'enfant (**Tableau 2**).

RECODAGE DES VARIABLES

Les variables quantitatives ont été "discrétisées" afin de mieux rendre compte d'éventuelles liaisons non linéaires avec le retard de taille. Lorsque le nombre initial de modalités d'une variable qualitative était trop important (information éclatée au maximum), nous avons décidé de procéder à un regroupement des modalités afin d'obtenir des effectifs équilibrés dans les classes (plus satisfaisant pour notre analyse statistique) tout en conservant une possibilité d'interprétation sur le plan épidémiologique. Les variables concernées par ces recodages sont signalées dans le **Tableau 2**.

REGROUPEMENT DE VARIABLES

Certaines variables explicatives retenues ont été regroupées sous forme d'indices synthétiques afin de limiter le nombre de variables à inclure dans les modèles sans toutefois perdre l'information qu'elles apportent. Il s'agit de 11 variables décrivant le logement et de 14 décrivant les biens que possède le ménage. Deux indices ont été construits : un indice de biens possédés par le ménage et un indice de niveau économique à partir de l'indice de biens, des variables décrivant le logement et de deux variables particulières décrivant la possession de moutons avant et après Tabaski[6] (**Tableau 2**).

Indice de biens :

Pour résumer la possession de biens du ménage et sachant que chaque bien n'a pas la même valeur, à défaut d'un relevé de prix plus précis, on pondère chaque variable par une pseudo valeur monétaire estimée (pondérations décrites dans le **Tableau 2**). Le calcul consiste à faire ensuite la somme pondérée de ces variables, la modalité correspondant à la possession du bien étant codée 1 et la non-possession codée 0. Plus cet indice est élevé plus le ménage possède des biens dont la valeur et/ou le nombre est (sont) élevé(s) La variable numérique obtenue a été ensuite découpée en quintiles et introduite dans le calcul de l'indice de niveau économique.

Indice de niveau économique : (Traissac et al., 1997)

L'indice est construit à partir d'une analyse factorielle des correspondances (AFC) effectuée à l'aide de la procédure CORRESP (SAS Institute Inc., 1989 b) dans SAS.

[6] Fête musulmane ("Aïd El Kebir" dans les pays du Maghreb) dont un des aspects est le sacrifice rituel d'un mouton. Du point de vue religieux mais aussi socio-culturel, c'est un des moments forts de l'année. Les familles y compris les plus pauvres investissent une partie très importante de leurs ressources (voir même s'endettent lourdement) pour le célébrer dignement (d'où l'importance de ces deux variables).

Tableau 3 : Distribution de l'échantillon selon un état de santé M dichotomique et un facteur de risque X dichotomique

	M^+	M^-	Total
X_1	a	c	n_1
X_0	b	d	n_0
Total	m_1	m_0	n

Tableaux 4a, 4b et 4b' : Distribution de l'échantillon selon un état de santé M dichotomique et un facteur de risque X à trois modalités X_0, X_1 et X_2

4a.

	M^+	M^-	Total
X_2	a_2	c_2	n_2
X_1	a_1	c_1	n_1
X_0	b	d	n_0
Total	m_1	m_0	n

4b'.

	M^+	M^-	Total
X_2	a_2	c_2	n_2
X_0	b	d	n_0
Total	$m_{1''}$	$m_{0''}$	n''

4b.

	M^+	M^-	Total
X_1	a_1	c_1	n_1
X_0	b	d	n_0
Total	$m_{1'}$	$m_{0'}$	n'

Tableaux 5a, 5b et 5c : Distribution de l'échantillon selon un état de santé M et un facteur de risque X dichotomiques et un tiers facteur T à trois modalités T_1, T_2 et T_3

- 5a. Modalité T_1

	M^+	M^-	Total
X_1	a_1	c_1	n_1
X_0	b_1	d_1	n_0
Total	$m_{1'}$	$m_{0'}$	n'

- 5b. Modalité T_2

	M^+	M^-	Total
X_1	a_2	c_2	n_2
X_0	b_2	d_2	n_0
Total	$m_{1''}$	$m_{0''}$	n''

- 5c. Modalité T_3

	M^+	M^-	Total
X_1	a_3	c_3	n_3
X_0	b_3	d_3	n_0
Total	$m_{1'''}$	$m_{0'''}$	n'''

L'idée est de chercher un axe d'inertie maximale (décrivant un gradient de niveau économique) du nuage des ménages pour ce qui concerne les 12 variables choisies. Une variable synthétique décrivant le niveau économique du ménage est ainsi créée à partir des coordonnées de l'axe retenu. Cet indice est alors découpé en terciles.

Malgré leur participation au calcul de cet indice, les variables "Source d'eau" et "Sanitaires" ont été conservées dans les analyses ultérieures car elles ne décrivent pas seulement le niveau économique du ménage mais aussi l'état de salubrité du milieu (**Tableau 2**).

B. LES INDICES D'ASSOCIATION : ODDS-RATIO (OR) ET RISQUE RELATIF (RR)

L'étude d'un facteur de risque consiste à analyser la répartition des sujets selon l'état de santé dans chaque modalité du facteur de risque. Les individus malades sont notés M^+ et les autres M^-. Concernant le facteur de risque X, plusieurs cas peuvent se présenter : ou bien X est une variable dichotomique (les deux catégories sont alors notées X_1 pour les sujets exposés et X_0 pour les sujets non exposés), ou bien X est une variable à plus de deux modalités, ou alors il existe plusieurs facteurs que l'on doit prendre en compte dans la mesure du risque.

1. Relation entre un facteur de risque dichotomique et un état de santé dichotomique (**Tableau 3**)

Le test du khi² permet de tester l'existence d'une association entre l'état de santé et le facteur de risque. Mais celui-ci ne suffit pas pour documenter la nature de cette association. En particulier il ne dit rien sur la force de liaison des deux variables. Il existe des mesures dites d'association qui permettent de quantifier l'intensité de cette liaison.

- Une mesure d'association "naturelle" est le **risque relatif brut,** noté RR_b et aussi appelé rapport de prévalences. C'est le rapport de probabilité P_1 d'être malade lorsqu'on est exposé au facteur de risque sur la probabilité P_0 d'être malade lorsqu'on n'est pas exposé au facteur.

$$RR_b = \frac{P_1}{P_0}$$ **Formule 2 : Risque relatif brut théorique**

RR_b est estimé sur l'*échantillon* par $\hat{R}R_b = \frac{a\ n_1}{b\ n_0}$

- Une autre mesure de l'association entre un facteur de risque et un état de santé est **l'odds-ratio brut,** noté OR_b :

$$OR_b = \frac{P_1\ (1-P_1)}{P_0\ (1-P_0)}$$ **Formule 3 : Odds-ratio brut théorique**

OR_b est estimé sur l'*échantillon* par $\hat{O}R_b = \frac{ad}{bc}$

Une probabilité de maladie identique dans les deux classes d'exposition indique une absence de lien entre la maladie et le facteur de risque. Cette absence d'association se traduit formellement par un risque relatif et un odds-ratio tous deux égaux à 1. Si cette probabilité de maladie est plus grande dans la classe d'exposition que celle de non-exposition, les indices d'association prennent des valeurs supérieures à 1 qui traduisent alors un risque augmenté. Le risque est diminué si celles-ci sont inférieures à 1.
Le risque relatif est plus facilement interprétable que l'odds-ratio : c'est le facteur par lequel le risque de maladie est multiplié en présence du facteur de risque. L'odds-ratio, lui, exprime aussi la relation entre P_0 et P_1 mais de façon moins immédiate. Il est néanmoins plus souvent utilisé que le risque relatif car celui-ci peut-être estimé dans tous les types d'enquête (transversale, cas-témoins, cohorte)[7]. L'odds-ratio est très souvent supérieur au risque relatif et, au-delà d'une prévalence de maladie supérieure à 20% (Bouyer, 1995), les deux quantités deviennent sensiblement différentes. Mais lorsque la prévalence de la maladie étudiée est faible et que le risque relatif n'est pas trop élevé, il est facile de montrer que OR_b est une bonne approximation de RR_b. En effet lorsque P_1 et P_0 sont petits (**Formule 3**), ces deux probabilités peuvent être négligées devant 1 et donc la formule de l'odds-ratio se simplifie en celle du risque relatif (**Formule 2**).
Même s'il existe des différences numériques entre ces deux mesures, celles-ci évoluent toujours dans le même sens. Néanmoins cette interprétation des OR comme approximations des RR dans le cas de "maladies rares" ne doit pas être faite abusivement lorsque les conditions ne sont pas vérifiées. C'est pourquoi nous allons nous intéresser à ces deux mesures et les comparer tout au long de notre analyse.

Les estimations ponctuelles des OR et RR permettent de raisonner au niveau de l'échantillon. Ces valeurs peuvent varier d'un échantillon à l'autre. Pour faire face à ces fluctuations d'échantillonnage et pouvoir extrapoler les résultats au niveau de la population[8], on a calculé les intervalles de confiance associés à ces estimations. On a ainsi de bonnes chances de trouver dans ces intervalles les valeurs réelles des OR et des RR. L'interprétation de ces intervalles de confiance est la suivante : s'ils ne contiennent pas la valeur 1 on peut conclure soit à l'augmentation du risque lié au facteur étudié (les bornes de l'intervalle de confiance doivent être strictement supérieures à 1) soit à la diminution du risque lié au facteur étudié (les bornes doivent être strictement inférieures à 1). Si ces intervalles de confiance contiennent la valeur 1 on ne peut conclure sur la significativité de OR_b et de RR_b à l'échelle de la population.

2. Facteur de risque à plus de deux classes

Soient un état de santé dichotomique et un facteur de risque à trois modalités (X_0, X_1 et X_2). OR_b et RR_b exprimant la relation entre le risque de maladie lorsqu'on est exposé au facteur et ce même risque lorsqu'on n'est pas exposé à ce même facteur, il faudra se replacer dans le cadre d'un facteur de risque dichotomique et donc choisir une classe de référence (sera la modalité présentant par exemple à la fois à priori le moins de risque pour le retard de taille et ayant un effectif pas trop petit). Si on choisit X_0 comme classe de référence, on aura deux OR_b et deux RR_b à calculer (**Tableaux 4a**, **4b** et **4b'**) : $OR_{b''}$ et $RR_{b''}$ comparant les classes X_2 et X_0 et, $OR_{b'}$ et $RR_{b'}$ comparant les classes X_1 et X_0. Pour un facteur de risque à k modalités on pourra calculer k-1 OR_b et RR_b.

[7] Pour plus de précisions, se référer à Bouyer et al. 1995
[8] Sous l'hypothèse que ces échantillons soient representatifs de cette population

3. Deux facteurs de risque

Plusieurs facteurs peuvent être considérés simultanément. La mesure du risque dépend alors non seulement de chaque facteur, mais aussi de la façon dont la présence d'un facteur modifie l'association de l'autre avec la maladie. On distingue deux cas :

- Facteurs modificateurs d'effet :

Si on s'intéresse à la liaison entre un facteur de risque X et une maladie M, un tiers facteur T est dit modificateur d'effet si la relation entre ce facteur de risque et la maladie n'est pas la même dans les strates définies par les différents niveaux de T. En théorie il suffit de comparer la mesure d'association (OR_b ou RR_b) entre X et M dans chaque strate de T (**Tableaux 5a, 5b** et **5c**). Ce dernier est modificateur d'effet si les mesures d'association sont significativement différentes d'une strate à l'autre. En pratique cette analyse désagrégée est réalisée par l'étude de modèles de régression trivariés où on teste l'interaction entre le facteur de risque et le tiers facteur T.

- Facteurs de confusion :

S'il n'est pas modificateur d'effet, un tiers facteur T est dit de confusion s'il explique, totalement ou partiellement, une association réelle entre X et M, ou si, au contraire, il masque totalement ou partiellement, cette association. Un facteur de confusion doit répondre aux trois caractéristiques suivantes : il doit être un facteur de risque de la maladie c'est-à-dire lié à M aussi bien dans la classe d'exposition du facteur de risque que dans la classe de non-exposition. Il doit être également lié à X. Enfin il ne doit pas être une étape intermédiaire dans le mécanisme causal liant X et M. Si T n'est pas modificateur d'effet, il est justifié de considérer un indice d'association "moyen" construit à partir des indices d'association observés dans chaque strate défini par T. Le tiers facteur T sera de confusion si l'indice brut et l'indice moyen (appelé aussi indice ajusté pour T) sont très différents. De la même manière que pour les facteurs modificateur d'effet, l'étude des facteurs de confusion peut aussi se faire par des modèles de régression trivariés où on compare l'effet brut du facteur de risque X et son effet ajusté au facteur de confusion T.

C. LES MODELES UTILISES

La variable à expliquer étant dichotomique, il n'est pas possible d'utiliser un modèle linéaire général puisque cela suppose que l'hypothèse de distribution faite sur la variable réponse soit normale. Cette contrainte est trop forte et difficilement envisageable pour la variable que l'on cherche à expliquer. De plus on veut pouvoir estimer des indices d'association (odds-ratios et risques relatifs) que le modèle linéaire général ne permet pas d'obtenir. Il faut donc trouver des modèles adaptés pour expliquer des variables réponses de type dichotomique et pour estimer des indices d'association. Les **modèles logistique et log-binomial** répondent à ces deux objectifs.

Ces deux modèles, ainsi que le modèle linéaire général, appartiennent à la classe des modèles linéaires généralisés, chacun de ces modèles étant caractérisé par une fonction de lien et une hypothèse de distribution sur la variable à expliquer.

Tableau 6 : Ajustement des intervalles de valeurs par la fonction de lien

Modèle	Variable expliquée Y		Combinaison des $X_i\beta_i$	Fonction de lien g	E(Y) prédite par le modèle
Linéaire général	Continue	$]-\infty ; +\infty[$	$]-\infty ; +\infty[$	Identité	$]-\infty ; +\infty[$
Logistique	Dichotomique	$[0 ; 1]$	$]-\infty ; +\infty[$	Logit	$[0 ; 1]$
Log-binomial	Dichotomique	$[0 ; 1]$	$]-\infty ; +\infty[$	Ln	$[0 ; +\infty[$

Tableau 7 : Quelques cas particuliers du modèle linéaire généralisé

Type de modèle	Variable expliquée Y	Distribution de Y	Fonction de lien g
Modèle linéaire général	Quantitative continue	normale	Identité
- Régression linéaire simple ou multiple*	Quantitative continue	normale	Identité
- Analyse de variance*	Quantitative continue	normale	Identité
- Analyse de covariance*	Quantitative continue	normale	Identité
Régression logistique	Dichotomique	binomiale	logit
Régression log-binomiale	Dichotomique	binomiale	ln
Régression log-poisson	Comptage	poisson	ln

Tableau 8 : Quelques exemples de valeurs de la fonction de variance V(μ)

Loi	V(μ)	ϕ
Normale	1	σ^2
Binomiale	$\mu(1-\mu)$	1
Poisson	μ	1

* Ce modèle est lui-même un cas particulier du modèle linéaire général

1. Notions sur le modèle linéaire généralisé (Dobson, 1990)

a) Fonction de lien

Soient Y une variable réponse quelconque et X_1, X_2,..., X_i,...,X_n ($1<i<n$) des variables explicatives (discrètes et/ou continues). L'espérance de Y, noté E(Y), est une fonction d'une combinaison linéaire des variables explicatives :

$$E(Y) = g(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + ... + \beta_n X_n)$$

avec β_i le paramètre estimé par le modèle et associé à la variable X_i, et g appelée fonction de lien

Dans le cas du modèle linéaire général, aucune contrainte n'est imposée sur les valeurs des β_i : Y étant continue, son espérance peut prendre n'importe quelle valeur. La fonction de lien g est ici la fonction identité I.

$$\underset{\alpha+\beta_1X_1+\beta_2X_2+\ .\ +\beta_iX_i+\ +\beta_nX_n}{]-\infty\ ;\ \infty[} \quad \underset{g^{-1}=I}{\overset{g=I}{\rightleftharpoons}} \quad \underset{E(Y)}{]-\infty\ ;\ \infty[}$$

En revanche dans le cas des modèles logistique et log-binomial, E(Y) doit nécessairement appartenir à l'intervalle [0 ; 1]. Il faut donc trouver une fonction g qui, à partir des $\sum X_i B_i$, permet de générer des espérances comprises dans cet intervalle.

$$\underset{\alpha+\beta_1X_1+\beta_2X_2+\ +\beta_iX_i+\ +\beta_nX_n}{]-\infty\ ;\ \infty[} \quad \underset{g'^{-1}}{\overset{g'}{\rightleftharpoons}} \quad \underset{E(Y)}{[0\ ;\ 1]}$$

La fonction de lien est donc définie comme une fonction permettant de mettre en bijection l'espace des valeurs de la variable réponse et celui des variables explicatives (**Tableau 6**). Un choix adéquat de cette fonction permet également d'interpréter les paramètres estimés par le modèle en terme d'indices d'association : la fonction de lien **logit** permettra d'interpréter les paramètres du modèle logistique en terme d'odds-ratios; la fonction **ln** permettra d'estimer des risques relatifs. Il n'est pas toujours possible d'obtenir une bonne fonction de lien tenant compte simultanément de ces deux critères. C'est le cas notamment de la fonction **ln** qui ne permet pas toujours d'obtenir des valeurs de E(Y) comprises entre 0 et 1 (**Tableau 6**). Nous verrons dans la suite comment utiliser ce modèle en contournant le problème.

b) Distribution de la variable réponse

Dans le modèle linéaire généralisé chaque distribution choisie pour modéliser la variation de Y_i autour de son espérance doit nécessairement appartenir à une famille de distribution appelée famille exponentielle dont les lois normale, binomiale et de poisson font partie (**Tableau 7**). Cette famille est caractérisée, entre autres, par une fonction reliant l'espérance de la distribution à sa variance. Si μ représente la moyenne de la variable Y, sa variance V(Y) peut être exprimée selon la formule :

$$V(Y) = \frac{V(\mu) * \phi}{\omega}.$$

Formule 4 : Lien entre moyenne et variance dans la théorie des modèles linéaires généralisés

où $V(\mu)$ est la fonction de variance, ϕ un paramètre de dispersion et ω le poids de chaque *observation.*

Les poids sont en général fixés à 1. Des exemples de valeurs de V(μ) sont données dans le **Tableau 8**. Pour les lois binomiales et de poisson il est possible dans certains cas d'estimer le paramètre ϕ.

c) Estimation des paramètres du modèle

L'estimation des paramètres se fait par la méthode du maximum de vraisemblance décrite dans de nombreux ouvrages de statistique (Dagnelie, 1973). Cette méthode a pour principe de choisir comme estimation de tout paramètre β la valeur la plus vraisemblable, c'est-à-dire celle qui donne à l'échantillon observé, la plus forte probabilité d'apparition. On définit la fonction de vraisemblance comme la probabilité ou la densité de probabilité relative aux valeurs observées y_1, y_2, ..., y_n, exprimée en fonction du ou des paramètres de la population Pour un échantillon aléatoire et simple dont les unités statistiques sont indépendantes et pour une population définie par un seul paramètre β, la fonction de vraisemblance s'écrit :

$$L = f(y_1;\beta) f(y_2;\beta) \ldots f(y_n;\beta)$$

Les estimations du maximum de vraisemblance correspondent par définition au maximum de cette fonction. La recherche de ce maximum peut être réalisée en annulant la dérivée de la fonction par rapport à β :

$$\frac{\partial Ln(L)}{\partial \beta} = 0$$

On aura autant de dérivées à calculer qu'il y aura de paramètres. Ces équations n'ont pas toujours de solutions exactes et donc en particulier dans nombre de cas de modèles linéaires généralisés la recherche de ces solutions se fait par l'utilisation de méthodes numériques. L'algorithme de recherche génère automatiquement des valeurs de départ pour chaque paramètre de la fonction de vraisemblance à déterminer. Une probabilité d'apparition de l'échantillon est estimée. Ces valeurs de départ sont ensuite itérées. A chaque itération une probabilité d'apparition est calculée et comparée à la précédente. Si L possède un maximum la méthode permet de converger vers le point où la probabilité est maximale.

Pour certaines fonctions de lien il arrive qu'à certaines itérations, les probabilités estimées soient supérieures à 1 (cas du modèle log-binomial). L'algorithme se bloque alors et la recherche de la fonction de vraisemblance s'arrête. Ce problème peut être évité en choisissant de bonnes valeurs de départ pour la recherche itérative de L. Des exemples de valeurs seront présentés un peu plus loin.

2. Estimation d'odds-ratios : le modèle logistique (Bouyer, 1991)

L'utilisation de la fonction de lien **logit** permet une interprétation des paramètres en terme de OR. Selon le nombre de facteurs inclus dans le modèle on pourra obtenir soit un odds-ratio brut (OR_b) soit un odds-ratio ajusté (OR_a) sur les autres facteurs. L'inclusion de termes d'interaction permettra également la gestion des facteurs modificateur d'effet.

a) Un seul facteur de risque

Soit $P(M^+/X)$ la probabilité d'avoir un retard de taille connaissant le facteur de risque X. Dans le cas d'un seul facteur X, l'expression mathématique du modèle logistique est la suivante : $P(M^+/X)=P=f(X)$ où f est la fonction logistique :

$$f(X) = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

Une forme plus simple est d'écrire le modèle logistique sous sa forme logit :

$$Logit(P) = \ln\frac{P}{1-P} = \alpha + \beta X$$ **Équation 1 : Modèle logistique pour un facteur de risque**

Si le facteur de risque X est codé 0/1 (non exposé/exposé), le modèle logistique s'écrit :

- Pour X=0

$$Logit\left[P\left(M^+/X_{=0}\right)\right] = Logit(P_0) = \ln\frac{P_0}{1-P_0} = \alpha$$

- Pour X=1

$$Logit\left[P\left(M^+/X_{=1}\right)\right] = Logit(P_1) = \ln\frac{P_1}{1-P_1} = \alpha + \beta X$$

Connaissant la formule de l'odds-ratio on peut écrire :

$$\ln(OR_b) = \ln\frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \ln\frac{P_1}{1-P_1} - \ln\frac{P_0}{1-P_0} = Logit(P_1) - Logit(P_0) = \beta X$$

Finalement on a :

$$\boxed{OR_b = \exp(\beta)}$$ **Formule 5 : Relation entre OR brut et le modèle logistique**

On pourra donc estimer directement OR_b à partir des estimations des paramètres du modèles par $\hat{OR}_b = \exp(\hat{\beta})$. De même un intervalle de confiance sur OR_b pourra être obtenu, basé sur la normalité asymptotique des estimateurs du maximum de vraisemblance $\hat{\beta}$ par

$\exp\left(\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}} ; \hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}}\right)$ $(p = 0.95)$ où $\hat{\sigma}_{\hat{\beta}}$ est l'écart-type d'estimation de $\hat{\beta}$ dans le cadre du modèle logistique. Si le facteur X a un nombre de modalités supérieur à 2, le choix de la catégorie de référence se fait par omission de la variable indicatrice correspondante dans le modèle, parmi les p codant le facteur.

b) Plusieurs facteurs de risque

Le modèle précédent se généralise facilement à plusieurs facteurs de risque comportant un nombre quelconque de modalités.

Soit un modèle à deux facteurs :

1) X à 2 modalités X_1 (X_2)
2) X' à 3 modalités X'_1, X'_2, (X'_3)

La modalité entre parenthèses représente la classe de référence.

$$P\left(M^+ / X, X'\right) = P = f(X) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_1' X'_1 + \beta_2' X'_2)}}$$

$$Logit(P) = \ln \frac{P}{1-P} = \alpha + \beta_1 X_1 + \beta_1' X'_1 + \beta_2' X'_2$$

Équation 2 : Modèle logistique à 2 facteurs de risque sans interaction[9]

Ici la valeur du coefficient β_1 dépend de la présence des autres variables. Autrement dit une même variable X n'aura pas le même coefficient dans un modèle où elle est seule présente et dans un modèle où figurent d'autres variables. Pour un facteur de risque donné on obtient un OR ajusté aux autres facteurs se trouvant dans le modèle. Deux cas peuvent se présenter :

- **Pas de facteur modificateur d'effet pour le facteur de risque mesuré**

L'odds-ratio du facteur de risque X (par exemple) ajusté aux autres facteurs se calcule selon la formule :

$$\hat{OR}_{a\left(X_1=1 \atop X_2=0\right)} = \exp(\hat{\beta}_1)$$

Formule 6 : Relation entre OR ajusté et le modèle logistique sans interaction

De même que précédemment un intervalle de confiance sera déduit directement des estimations des paramètres du modèle.

- **Présence d'un facteur modificateur d'effet pour le facteur de risque mesuré**

Soit un modèle à deux facteurs avec interaction :

1) X à 2 modalités X_1 (X_2)
2) X' à 3 modalités X'_1, X'_2, (X'_3)
3) Interaction XX' $X_1X'_1$ $X_1X'_2$ $X_1(X'_3)$

[9] En l'absence de facteurs modificateurs d'effet

$(X_2)X'_1$ $(X_2)X'_2$ $(X_2)(X'_3)$

On considère que le facteur X' est modificateur d'effet pour X. Les paramètres des catégories de référence n'étant pas estimés, le modèle logistique avec interaction s'écrit :

$$Logit(P) = \ln\frac{P}{1-P} = \alpha + \beta_1 X_1 + \beta_1' X'_1 + \beta_2' X'_2 + \gamma_{11} X_1 X'_1 + \gamma_{12} X_1 X'_2$$

Équation 3 : Modèle logistique à 2 facteurs de risque avec interaction

Le calcul de l'odds-ratio de X ajusté sur X' est corrigé par un terme d'interaction spécifique à chaque strate du facteur modificateur d'effet. On aura autant de termes correcteurs que de strates de X', un des termes étant toujours nul.

Formules 7a, 7b et 7c : Relation entre OR ajusté et le modèle logistique avec interaction

$$\hat{OR}_{a\left(X_1=1,\ X_2=0\right)} = \exp(\hat{\beta}_1 + \hat{\gamma}_{11})$$ pour la strate 1 du facteur X'

$$\hat{OR}_{a\left(X_1=1,\ X_2=0\right)} = \exp(\hat{\beta}_1 + \hat{\gamma}_{12})$$ pour la strate 2 du facteur X'

$$\hat{OR}_{a\left(X_1=1,\ X_2=0\right)} = \exp\left(\hat{\beta}_1 + \underbrace{\hat{\gamma}_{13}}_{=0}\right)$$ pour la strate 3 du facteur X'

La méthode de calcul peut se généraliser à des facteurs de risque et modificateurs d'effet ayant un nombre de modalités quelconque. Comme précédemment l'estimation et le calcul d'un intervalle de confiance pour l'indice d'association se déduisent des estimations des paramètres du modèle. Néanmoins dans la mesure où on a une somme de paramètres, il est nécessaire de tenir compte de la covariance entre les estimateurs. L'**Annexe 12** présente un exemple de calcul pour un facteur de risque et un facteur modificateur d'effet tous deux à trois modalités.

3. Estimation de risques relatifs : le modèle log-binomial

L'hypothèse faite sur la distribution du retard de taille (loi binomiale) est la même que pour le modèle logistique. Cependant la fonction de lien n'est plus la fonction **logit** mais la fonction **ln** dont les paramètres s'interprètent en terme de RR. Cette fonction s'écrit :

- **Pour un facteur de risque :**

$$f(X) = P = e^{\alpha+\beta X}$$

Sa forme logarithmique est :

$$\ln(P) = \alpha + \beta X$$

Équation 4 : Modèle log-binomial pour un facteur de risque

De la même manière que pour OR_b :

$$\ln(P_1) = \alpha + \beta X \quad \text{et} \quad \ln(P_0) = \alpha$$

$$\ln(RR_b) = \ln\frac{P_1}{P_0} = \beta X$$

Le risque relatif brut s'écrit donc :

$$\boxed{RR_b = \exp(\beta)}$$

Formule 8 : Relation entre RR brut et le modèle log-binomial

L'estimation de RR_b et de son intervalle de confiance pourra se faire, comme pour OR_b, à partir des estimations des paramètres du modèle.

- **Pour plusieurs facteurs de risque**

Les formules sont toutes identiques au modèle logistique. Les OR sont remplacés par des RR.

$$P\left(M^+ / X, X'\right) = P = f(X) = e^{\alpha+\beta_1 X_1+\beta_1' X'_1+\beta_2' X'_2}$$

$$Ln(P) = \alpha + \beta_1 X_1 + \beta_1' X'_1 + \beta_2' X'_2$$

Équation 5 : Modèle log-binomial à 2 facteurs de risque sans interaction

$$Ln(P) = \alpha + \beta_1 X_1 + \beta_1' X'_1 + \beta_2' X'_2 + \gamma_{11} X_1 X'_1 + \gamma_{12} X_1 X'_2$$

Équation 6 : Modèle log-binomial à 2 facteurs de risque avec interaction

4. Tests dans les modèles

a) Tests de rapport de vraisemblance

Quel que soit le type de modèle utilisé (logistique ou log-binomial), cette méthode permet de comparer à 0 plusieurs coefficients à la fois et notamment de tester la signification des facteurs modificateurs d'effet et des facteurs de confusion. Elle consiste à comparer les vraisemblances de deux modèles dits "emboîtés", l'un étant un cas particulier de l'autre.
Soient :

- un modèle 1 à p paramètres ($\alpha, \beta_1, \ldots, \beta_p$) et de vraisemblance L_1
- un modèle 2 à q paramètres ($\alpha, \beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_q$) et de vraisemblance L_2

avec $p<q$

Les hypothèses testées sont :

H_0 : le modèle 2 n'est pas meilleur que le 1 ($\beta_{p+1}=\beta_{p+2}= \ldots=\beta_q=0$)

H_1 : le modèle 2 est meilleur que le 1

On calcule le rapport k que l'on compare à 1 :

$$k = L_2 / L_1$$

Si k>1 L_2 est meilleur
Si k≅1 L_2 n'est pas meilleur

On s'intéresse en fait à :

$$2 \ln k = 2 \ln(L_2) - 2 \ln(L_1)$$

Sous des conditions assez générales, 2 ln k suit une loi du khi² à (q-p) ddl. De cette propriété est déduite le test du rapport de vraisemblance.

- **Détermination des facteurs modificateurs d'effet**

Rechercher un facteur modificateur d'effet consiste à modéliser une interaction entre deux facteurs. La comparaison des deux modèles permet de tester la significativité de cette interaction (**Équations 7a** et **7b**). Soient un facteur de risque X et un facteur modificateur d'effet X' tous deux dichotomiques. Les termes X_1 et X'_1 dans les équations ci-dessous ne sont pas les facteurs mais les indicatrices de ces facteurs. Si les facteurs ne sont pas dichotomiques, il faudra plusieurs indicatrices pour coder chaque facteur.

Les hypothèses testées sont :

H_0 : $\gamma_{12,12}=0$ (le facteur X' n'est pas modificateur d'effet)
H_1 : $\gamma_{12,12}\neq 0$ (le facteur X' est modificateur d'effet)

Équations 7a et 7b : Les modèles comparés dans la recherche des facteurs modificateurs d'effet

	Modèle 1	**Modèle 2**
Logistique	Logit P = $\alpha_1 + \beta_{11}X_1 - \beta'_{11}X'_1$	Logit P = $\alpha_2 + \beta_{12}X_1 + \beta'_{12}X'_1 + \gamma_{12,12}\mathbf{X_1X'_1}$
Log-binomial	Ln P = $\alpha_1 + \beta_{11}X_1 - \beta'_{11}X'_1$	Ln P = $\alpha_2 + \beta_{12}X_1 + \beta'_{12}X'_1 + \gamma_{12,12}\mathbf{X_1X'_1}$

Les indices ij correspondent à i . paramètre de la modalité X_i ou X'_i
j : type de modèle (1 ou 2)

Les paramètres d'une même variable ne sont pas écrits avec les mêmes symboles dans les deux modèles (β_{12} et β'_{12} au lieu de β_{11} et β'_{11}) car la présence de l'interaction $X_1X'_1$ modifie la signification et les valeurs des coefficient de X_1 et de X'_1.

Les estimations des coefficients α_1, β_{11} et β'_{11}(et α_2, β_{12}, β'_{12} et $\gamma_{12,12}$) sont celles pour lesquelles la vraisemblance L de chaque modèle est maximale. Le test du rapport de vraisemblance compare L_1 et L_2, les valeurs de ces vraisemblances maximales. Si H_0 est vraie, 2 ln k suit une loi de khi² à 1 ddl.

- **Recherche des facteurs de confusion**

Deux comparaisons sont réalisées : la première teste l'effet brut du facteur X et la deuxième l'effet de X ajusté à X' (**Équations 8a et 8b**). Les hypothèses à tester sont :

- Test 1 : $\mathbf{H_0}$ **:** $\beta_{121} = 0$ (pas d'effet brut de X_1)
  $\mathbf{H_1}$ **:** $\beta_{121} \neq 0$ (effet brut de X_1 significatif)
- Test 2 : $\mathbf{H_0}$ **:** $\beta_{122} = 0$ (pas d'effet de X_1 ajusté à X_2)
  $\mathbf{H_1}$ **:** $\beta_{122} \neq 0$ (effet de X_1 ajusté à X_2 significatif)

X' est un facteur de confusion dans le cas où on rejette H_0 du Test 1 et on conserve H_0 du Test 2. Le fait que l'effet brut de X soit significatif alors que son effet ajusté à X' ne l'est pas signifie que X' fait disparaître l'association entre le retard de taille et X et donc que X' est un facteur de confusion.

Équations 8a et 8b : Les modèles comparés dans les tests de confusion

	Modèle 1	**Modèle 2**
Logistique	Logit $P = \alpha_{111}$	Logit $P = \alpha_{121} + \beta_{121}\mathbf{X_1}$
	Logit $P = \alpha_{112} + \beta_{112}'X'_1$	Logit $P = \alpha_{122} + \beta_{122}\mathbf{X_1} + \beta_{122}'X'_1$
Log-binomial	Ln $P = \alpha_{111}$	Ln $P = \alpha_{121} + \beta_{121}\mathbf{X_1}$
	Ln $P = \alpha_{112} + \beta_{112}'X'_1$	Ln $P = \alpha_{122} + \beta_{122}\mathbf{X_1} + \beta_{122}'X'_1$

Les indices ijk correspondent à i : paramètre de la modalité X_i ou X'_i
j : type de modèle (1 ou 2)
k : type de Test (1 ou 2)

Les rapports de vraisemblance réalisés suivent aussi une loi de khi² à 1 ddl, dans le cas où les hypothèses H_0 sont vraies.

b) Validité des modèles

Deux grandes questions que l'on doit se poser lorsqu'on étudie la validité d'un modèle sont :

- Les conditions d'application du modèle sont-elles réunies ?
- Le modèle s'ajuste t-il aux données avec une précision suffisante ?

Dans le cadre du modèle linéaire généralisé une façon de répondre à ces deux questions est l'étude de la déviance (Dobson, 1990).

La **déviance** D est définie, à une constante près, comme la différence de vraisemblance entre le modèle « maximal ou saturé » et le modèle étudié. Le **modèle maximal**, qui décrit les données de manière complète, est un modèle linéaire généralisé ayant la même distribution que le modèle étudié (binomiale), la même fonction de lien (**logit** ou **ln**) et autant de paramètres que le nombre d'observations de l'échantillon étudié (N=4477). Pour un même jeu de données la définition du modèle maximal n'est pas unique : elle dépend entre autres du

mode de présentation des données (individuelles ou regroupées) ainsi que du nombre de variables prises en compte pour la définition (Simonoff, 1998).

Sous certaines hypothèses on peut approximer la distribution de la déviance d'un modèle à p paramètres par un khi² à N-p ddl. Néanmoins, souvent cette approximation n'est pas très bonne sauf pour le cas d'un modèle linéaire général pour lequel le résultat est exact et donc on ne construit pas formellement un test d'adéquation basé sur cette distribution. Toutefois l'adéquation du modèle aux données peut s'évaluer grossièrement en comparant la déviance estimée à partir des données à une distribution du khi² appropriée. Si le modèle est bon on peut espérer une valeur de D proche du milieu de la distribution. Ceci est facile à évaluer puisque l'espérance d'une variable aléatoire distribuée selon une loi du khi² n'est autre que sa moyenne. Ainsi si un modèle à p paramètres décrit bien les données issues de l'échantillon étudié, D suit une loi du khi² à N-p ddl et :

$$\mathbf{D \cong N - p}$$

Lorsque le modèle s'ajuste mal aux données, on peut s'interroger sur l'hypothèse de distribution faite sur le retard de taille. Un phénomène courant lorsqu'on utilise des modèles basés sur la loi binomiale (en particulier la régression logistique et la régression log-binomiale) est la **surdispersion** : dans les données observées la variabilité des Y_i autour de leur espérance est supérieure à celle qui devrait être observée si les Y_i suivaient effectivement une loi binomiale. Autrement dit la variabilité des Y_i dans le modèle est inférieure à celle observée dans les données. Cela conduit à des tests trop puissants et des intervalles de confiance erronément trop étroits, dans la mesure où ils sont construits sous l'hypothèse de la loi binomiale. Par conséquent on peut conclure à tort sur la significativité des facteurs inclus dans le modèle. Le problème de la surdispersion peut se résoudre de deux manières :

- On change l'hypothèse de distribution faite sur le retard de taille (on pourra prendre par exemple la loi de poisson).

- On modifie la variance exprimée par le modèle (**Formule 4**) en estimant le paramètre de dispersion ϕ de sorte que la déviance se rapproche de N-p. Cette modification n'affecte en rien les estimations des paramètres du modèle. En revanche les estimations des écarts types de ces paramètres sont ajustées de manière appropriée (la variance des Y_i estimée par le modèle augmente).

Comme évoqué ci-dessus il n'est pas toujours possible d'obtenir une bonne estimation de la déviance. C'est notamment le cas lorsque l'on travaille sur des données individuelles ou des données regroupées à effectifs marginaux faibles. Le rapport $D/(N-p)$ n'est alors plus interprétable et on peut avoir recours au test d'ajustement proposé par Hosmer et Lemeshow (1989) dont le principe est le suivant : les sujets sont divisés approximativement en 10 groupes d'effectif à peu près égal basés sur les déciles des probabilités estimées d'apparition du retard de taille. Un test du khi² à 8 ddl compare dans chaque groupe le nombre d'observations observées et prédites par le modèle.

D. ELABORATION DES MODELES MULTIVARIES

Partant d'une liste de facteurs de risque, de facteurs de confusion et de facteurs modificateur d'effet retenus par les analyses précédentes, l'objectif général est de parvenir à un modèle décrivant au mieux le retard de taille tout en contenant le plus petit nombre possible de variables. On y gagne en précision sur les estimateurs des paramètres du modèle et on facilite l'interprétation des résultats (Bouyer et al., 1995). Les deux objectifs *diminution du nombre de variables et prise en compte des effets de confusion* peuvent être contradictoires : on donnera alors priorité au second.

La procédure d'élimination des variables utilisée est la méthode de régression pas à pas descendante. On teste de cette manière si des variables peuvent être supprimées. A chaque pas une variable est exclue du modèle avec un seuil de sortie fixé à l'avance. Le processus se poursuit jusqu'à ce qu'aucune variable ne remplisse les conditions de sortie dans le modèle. Pour des raisons à la fois technique (pas de procédure pas à pas pour le modèle linéaire généralisé dans SAS) et de prise en compte de confusion résiduelle (Rothman, 1986) cette procédure a été appliquée uniquement sur les interactions. On a donc cherché à retenir un minimum de termes d'interaction car leur présence rend l'interprétation du modèle délicate. En effet si l'on considère une interaction entre deux variables X_1 et X_2, on ne peut plus parler de l'OR (resp. RR) associé à X_1 ou à X_2. On a un OR (resp. RR) associé à X_1 pour un niveau de X_2 et un OR (resp. RR) associé à X_1 pour l'autre niveau de X_2 si on considère X_2 comme variable dichotomique. Les termes d'interaction d'OR (resp. RR) d'ordre supérieur à 2 n'ont pas été pris en compte.

A partir des paramètres estimés par les modèles on pourra calculer des mesures d'association ajustées à l'ensemble des facteurs (de risque, de confusion et modificateur d'effet).

E. MISE EN ŒUVRE INFORMATIQUE

GENMOD (SAS Institute Inc., 1997) est la procédure SAS gérant les modèles linéaires généralisés. Selon la fonction de lien choisie (logit ou log), cette procédure permet de générer des modèles logistiques et log-binomiaux. Elle a été utilisée à tous les stades de l'analyse (**Annexe 14**).

1. Les macros

(Sas Institute Inc., 1989a, 1989b, 1990a, 1990b, 1997)

- **Calcul et comparaison d'Odds-ratios et de Risques relatifs bruts (macro "%or_rr_brut ") :** (programme SAS voir **Annexe 15**)

Calculer une mesure d'association brute consiste à écrire un modèle de régression trivarié (logistique ou log-binomial) entre la variable à expliquer (le retard de taille) et un facteur de risque donné (la syntaxe SAS utilisée est précisée à l'**Annexe 10**). Pour calculer les indices d'association des 38 variables de notre étude, il aurait fallu écrire 38*2=76 modèles de régression. A partir des paramètres estimés par chaque modèle (Voir **Annexe 11 Sortie GENMOD Classique**, rubrique "Analysis of Parameter Estimates"), il faut ensuite calculer

un indice d'association et son intervalle de confiance. La macro SAS que j'ai écrite permet de compenser cette lourde tâche répétitive. En précisant l'état de santé étudié et une liste de facteurs de risque à nombre de modalités quelconques, la macro génère deux modèles bivariés (logistique et log-binomial) pour chaque combinaison état de santé / facteur de risque. Elle calcule ensuite pour chaque combinaison un OR et un RR brut ainsi que les intervalles de confiance associés. Les résultats sont enfin rassemblés dans un tableau permettant de comparer visuellement OR et RR pour un facteur de risque donné. Pour faciliter l'interprétation des résultats, le tableau fournit également les prévalences de retard de taille et les effectifs des modalités de chaque facteur de risque. Un exemple de sortie de cette macro est présenté à l'**Annexe 11.**

- **Détermination des facteurs modificateur d'effet (macro "%interac2")** (programme SAS voir **Annexe 16**)

A partir de modèles de régression trivariés logistique ou log-binomial, on teste l'interaction entre deux facteurs de risque. La variable à expliquer reste la même (cf. **Annexe 10** pour la syntaxe SAS utilisée). Le résultat du test du terme d'interaction dans le modèle est fourni dans la sortie GENMOD rubrique "LR Statistics for Type 3 Analysis" **(Annexe 11)**.

La recherche d'un (des) facteur(s) modificateur d'effet pour un facteur de risque à partir d'une liste variables consiste à tester toutes les interactions 2 à 2. Pour une liste de 38 variables par exemple, cela ferait 703 interactions à écrire et à analyser. La macro que j'ai écrite génère des modèles de régression trivariés pour chaque combinaison facteur de risque / facteur modificateur d'effet et rassemble dans un tableau le résultat de chaque test effectué (**Annexe 11**).

- **Recherche des facteurs de confusion (macro "%comparor")** (programme SAS voir **Annexe 17**)

La syntaxe utilisée pour la recherche d'un facteur de confusion est présentée à l'**Annexe 10**. Les résultats des Tests 1 et 2 du facteur modificateur d'effet (**C.4.a)**) sont fournis dans les rubriques "LR Statistics for Type 1 Analysis" et "LR Statistics for Type 3 Analysis" (**Annexe 11**). Le coté répétitif des calculs a nécessité l'écriture d'une macro. Celle-ci récapitule dans un tableau les résultats des tests, complétés par une comparaison visuelle de la mesure de risque brut et ajusté à un facteur (**Annexe 11**). Cette macro peut rechercher les facteurs de confusion en utilisant des modèles de régression trivariés logistique ou log-binomial.

2. Les problèmes de convergence liés au modèle log-binomial

Nous avons vu précédemment les problèmes que pouvaient occasionner la recherche de la fonction de vraisemblance maximale lorsque la fonction de lien ln était utilisée. Nous avons également précisé que ce problème pouvait être évité en choisissant de bonnes valeurs de départ pour la recherche itérative de cette fonction.

Les paramètres estimés par le modèle logistique peuvent être utilisés comme valeurs de départ pour la recherche de la fonction de vraisemblance du modèle log-binomial. Pour que les paramètres d'un modèle soient transposables à l'autre, les variables prises en compte doivent être identiques dans les deux modèles.

Il arrive parfois que l'utilisation des paramètres estimés par le modèle logistique se révèle inefficace pour la recherche de la meilleure fonction de vraisemblance Schouten et al. (1993)

propose dans ce cas de modifier les données de l'échantillon. La modification consiste à dupliquer les individus malades en individus non malades. Le nouvel échantillon est alors constitué de trois groupes : les malades, les non malades originaux et les nouveaux non malades. Il suggère ensuite d'appliquer un modèle logistique sur ce nouvel échantillon et d'utiliser les paramètres obtenus comme valeurs de départ pour la recherche de la fonction de vraisemblance du modèle log-binomial.

3. La programmation du test de Hosmer et Lemeshow
(programme SAS voir **Annexe 18**)

Ce test avait été utilisé la première fois par Hosmer et Lemeshow pour évaluer la validité du modèle logistique. La version 6.11 du logiciel SAS permet de faire en option ce test dans la procédure LOGISTIC (SAS Institute Inc., 1997). SAS propose uniquement ce test pour le modèle logistique. Néanmoins le test de Hosmer et Lemeshow n'est pas spécifique au modèle logistique. En effet celui-ci compare par un test du khi² des probabilités observées dans les données à celles prédites par le modèle étudié. Ce test peut donc être généralisable à n'importe quel autre modèle. C'est dans cet esprit qu'on a programmé ce test afin d'évaluer la validité de n'importe quel modèle linéaire généralisé et notamment celle du modèle logistique et du modèle log-binomial. On s'est basé pour cela sur les recommandations faites par ces auteurs (1989).

RESULTATS

Tableau 9 : Prévalences du retard de taille selon les indices de biens et de niveau économique

		n		Prévalences du retard de taille %	Liaison avec le retard de taille (Test du khi²)
Indice de biens (quintiles)	1	873	19%	16.8	
	2	847	19%	15.7	Khi²=39.0
	3	924	21%	13.1	P<0.01
	4	856	19%	11.3	
	5	977	22%	8.5	
Indice de niveau économique (terciles)	Elevé	1601	36%	10.2	Khi²=34.6
	Moyen	1511	34%	12.1	P<0.01
	Faible	1365	30%	17.2	

Une vue d'ensemble des résultats est présentée à l'**Annexe 19** et également sur la feuille volante distribuée avec ce rapport.

A. LES INDICES DE BIENS ET DE NIVEAU ECONOMIQUE

- Indice de biens :

L'indice de biens a été calculé à partir de 14 variables (**Tableau 2**). D'après les prévalences du retard de taille obtenues au **Tableau 9**, plus le ménage possède des biens dont la valeur et/ou le nombre est (sont) élevé(s), plus la prévalence du retard de taille est faible.

- L'indice de niveau économique

L'AFC a été réalisée à partir de 12 variables (**Tableau 2**). Les résultats sont présentés à l'**Annexe 1**. Le premier plan principal explique environ 28% de l'inertie totale. Les contributions des variables pour le premier axe sont assez élevées comparativement aux autres axes (15% pour la variable électricité et 20% pour l'indice de biens). Les coordonnées des modalités de chaque variable ont été reportées sur le graphique de cette même annexe. Pour des raisons de lisibilité les points situés au centre n'ont pas été représentés. L'analyse du graphique laisse apparaître clairement l'axe 1 comme un gradient opposant les ménages possédant le plus de biens et les meilleures conditions de logement aux ménages les plus défavorisés. On retient donc cet axe comme échelle économique synthétique qu'on découpe en terciles correspondant à un niveau de richesse décroissant (**Tableau 9**). Cette nouvelle variable qualitative ordinale sera utilisée dans les analyses ultérieures.

B. SELECTION DES VARIABLES ET DES UNITES STATISTIQUES

A l'issue de recodages et regroupements sur les variables initialement retenues comme facteurs de risque, 38 variables (**Tableau 2**) ont été testées pour leur liaison avec le retard de taille par test du Khi² sur les 4567 enfants de l'échantillon représentatif. L'**Annexe 2** présente les résultats en spécifiant les variables significatives au seuil α=20%. Ce seuil a été choisi car la règle qui conduit à ne retenir pour un modèle multivarié que les variables liées significativement au seuil 5% est trop stricte. On risque ainsi de laisser des effets de confusion résiduelle. En effet prendre un risque de 1ère espèce faible (probabilité de se tromper lorsqu'on affirme que le facteur influe sur le retard de taille) augmente celui de 2ème espèce (probabilité de se tromper lorsqu'on affirme que le facteur n'influe pas). Parmi les 25 variables liées significativement au retard de taille, 16 ont un seuil de signification inférieur ou égal à 5%.

L'analyse multivariée nécessite de disposer d'un échantillon où les unités statistiques n'ont aucune valeur manquante, simultanément, pour toutes les variables à entrer dans le modèle. Il a fallu supprimer 90 enfants de l'échantillon représentatif issu de l'enquête. Cet échantillon d'origine présente des valeurs manquantes pour seulement 3 des variables (indice de niveau économique, taille de la mère et traitement de la dernière diarrhée par pain de singe) parmi les 25 retenues pour l'analyse multivariée. Bien que cela concerne un pourcentage très faible d'enfants (2%) nous avons jugé nécessaire de vérifier l'absence de biais dans l'échantillon de 4477 enfants obtenu après exclusion des 90 unités statistiques par rapport à l'échantillon

représentatif d'origine. Pour cela nous avons comparé au seuil 5% la distribution des enfants entre les deux échantillons pour les 25 variables explicatives et le retard de taille : l'échantillon sans valeurs manquantes (n=4477) et l'échantillon d'enfants exclus (n=90 sauf pour l'indice économique où n=15, la parenté de l'enfant où n=76 et la taille de la mère où n=89).

La prévalence du retard de taille est la même dans les deux groupes (**Annexe 3**). Sept des 25 variables présentent une différence significative de distribution : on observe dans le deuxième échantillon (n=90) plus d'enfants dont le ménage dispose d'eau à volonté, de sanitaires privés, plus d'enfants dont la mère est épouse du chef de ménage ou elle-même chef de ménage, plus d'enfants dont la mère perçoit des allocations familiales, moins d'enfants dont la mère est d'ethnie Wolof, moins d'enfants dont le chef de ménage travaille dans l'informel, et moins d'enfants fils ou fille du chef de ménage.

On a enfin vérifié que sur l'échantillon de 4477 enfants, les 25 variables retenues précédemment étaient toujours liées au retard de taille au seuil α=20%.

On accepte le biais de distribution et on poursuit l'étude avec n=4477.

C. ODDS-RATIOS ET RISQUES RELATIFS BRUTS

Les indices d'association bruts ont été calculés pour les 25 variables liées au retard de taille. L'**Annexe 4** présente les OR_b et RR_b pour chacune de ces variables dont les résultats ont été regroupés en 4 sous tableaux (**4a** à **4d**) : le ménage, le chef de ménage, la mère et l'enfant. Bien que ces variables aient été retenues au seuil de 20%, les intervalles de confiance associées à ces mesures ont été calculés au seuil classique de 5%.

Il faut noter que ces indices d'association nous renseignent sur l'effet brut de chaque facteur sur le retard de taille, indépendamment des autres facteurs. De plus ces mesures sont de bons éléments de comparaison qui permettront à partir des mesures ajustées de juger de l'erreur que l'on peut faire en calculant un indice d'association entre le retard de taille et une exposition sans tenir compte des autres facteurs.

Les prévalences du retard de taille ont été calculées par modalité de chaque variable. Les valeurs extrêmes varient entre 6% chez les enfants de moins de 12 mois et 19% chez les enfants dont les mères font moins de 1.60 m (la prévalence moyenne est de 13%).
Par rapport aux classes de référence qu'on s'est fixées, on observe les risques de retard de taille les plus élevés pour des facteurs biologiques tels que l'âge de l'enfant et la taille de la mère et, d'autres facteurs tels que le niveau scolaire du chef de ménage et l'indice de niveau économique. Un enfant âgé de 36 à 47 mois a environ 3 fois plus de chances d'avoir un retard de taille que celui qui a moins de 12 mois. Un enfant dont la mère fait moins de 1.60 m a environ 2 fois plus de chances d'avoir un retard de taille que celui dont la mère fait plus de 1.65 m. Un enfant élevé par un chef de ménage non scolarisé a environ 2 fois plus de chances d'avoir un retard de taille que celui élevé par un chef de ménage possédant un brevet ou plus. Le risque de retard de taille est plus élevé pour les enfants des ménages avec un indice de niveau économique faible par rapport aux ménages ayant un indice élevé (le risque est multiplié par 1.7).
On a trois facteurs pour lesquels le risque de retard de taille est diminué par rapport aux classes de référence choisies à priori comme ayant les prévalences les plus faibles. Un enfant dont le chef de ménage réside depuis moins de 6 ans à Pikine a moins de chances d'avoir un retard de taille que celui dont le chef de ménage y réside depuis plus de 13 ans (risque multiplié par 0.7). Un enfant dont le chef de ménage travaille pour l'Etat a moins de chances d'avoir un retard de taille que celui dont le chef de ménage travaille dans le privé (risque

Tableau 10 : Liste des facteurs retenus pour l'analyse multivariée

Facteurs de risque potentiels et de confusion potentiels (25)
Densité d'individus par pièce
Source d'eau
Quantité d'eau disponible
Sanitaires
Indice de niveau économique
Sexe du chef de ménage
Age du chef de ménage
Ethnie du chef de ménage
Niveau scolaire du chef de ménage
Secteur d'activité du chef de ménage
Durée de résidence du chef de ménage à Pikine
Taille de la mère
Ethnie de la mère
Niveau scolaire de la mère
Situation matrimoniale de la mère
Parenté de la mère avec le chef de ménage
Enfants hospitalisés pour malnutrition grave
Allocations familiales
Sexe de l'enfant
Age de l'enfant
Connaissance du poids à la naissance
Parenté de l'enfant avec le chef de ménage
Dernière diarrhée récente
Dernière diarrhée traitée avec pain de singe
Existence d'un carnet de vaccination
Interactions (11)
Age de l'enfant * Sexe de l'enfant
Age de l'enfant * Taille de la mère
Connaissance du poids à la naissance * Parenté de l'enfant avec le chef de ménage
Source d'eau * Niveau scolaire de la mère
Connaissance du poids à la naissance * Parenté de la mère avec le chef de ménage
Densité d'individus par pièce * Indice de niveau économique
Dernière diarrhée récente * Densité d'individus par pièce
Quantité d'eau disponible * Allocations familiales
Indice de niveau économique * Secteur d'activité du chef de ménage
Sanitaires * Niveau scolaire de la mère
Dernière diarrhée récente * Sanitaires

Remarque : Les interactions en gras sont celles retenues à 5% par la procédure pas à pas

multiplié par 0.7). Le risque de retard de taille est plus faible chez les enfants dont la dernière diarrhée n'a pas été traitée avec du pain de singe (risque multiplié par 0.8).

D. LES FACTEURS D'AJUSTEMENT

Tous les tests de rapport de vraisemblance ont été réalisés avec les deux modèles (logistique et log-binomial) au seuil de signification α=5%. Les **Annexe 5** et **6** présentent les résultats des tests effectués sur les hypothèses nulles (cf. Sujets et Méthodes "tests basés sur la déviance" paragraphe **C.4.a**). Pour chaque facteur de risque , on a calculé les mesures d'association avec le retard de taille, ajustées individuellement à un facteur modificateur d'effet (**Annexe 8**) ou un facteur de confusion(**Annexe 9**). Une comparaison de ces mesures sera faite ultérieurement avec les résultats de l'analyse multivariée.

Les deux modèles fournissent globalement les mêmes résultats. La recherche des facteurs d'ajustement peut donc se faire par l'utilisation d'un seul modèle (la régression logistique par exemple).

1. Les facteurs modificateurs d'effet

Lorsqu'on teste une interaction, les modèles utilisés ne permettent pas de faire la distinction entre le facteur modificateur d'effet et le facteur de risque, les deux éventualités étant statistiquement possibles. Sur le plan épidémiologique une réflexion a été menée sur le choix des facteurs retenus comme modificateurs d'effet. Nous avons conservé uniquement 11 facteurs pour lesquels une hypothèse pouvait être émise.

La significativité des facteurs modificateur d'effet a été testée de manière individuelle (interaction d'ordre 2). Si deux facteurs sont modificateur d'effet pour un même facteur de risque, on ne peut rien dire sur leur significativité simultanée.

L'examen des mesures d'association dans chaque strate du facteur modificateur d'effet est en accord avec les tests de rapport de vraisemblance.

2. Les facteurs de confusion

En comparant les tests 1 et 2, on a pu dégager sept facteurs de confusion, deux d'entre eux étant de confusion pour plusieurs autres facteurs. Le test d'indépendance du khi² révèle une forte liaison entre chacun de ces facteurs et le retard de taille ($P<0.01$)

E. L'ANALYSE MULTIVARIEE

- **La régression pas à pas**

Réalisée sur la liste de facteurs obtenue par les analyses précédentes, elle a permis de retenir 6 interactions sur 11 étudiées (**Tableau 10**). Les modèle finaux contiennent donc ces 6 interactions plus les 25 variables non traitées par procédure pas à pas (prise en compte de la confusion résiduelle). Le seuil de sortie de ces interactions était de 5%. Les résultats des OR et des RR sont présentés de la même manière que pour les mesures d'association brutes (**Annexe 7a** à **7d**). La comparaison OR / RR est toujours possible. Une colonne supplémentaire a été ajoutée dans les tableaux et précise les éventuels ajustements sur les facteurs modificateurs d'effet.

Les indices d'association de chaque facteur avec le retard de taille ont été ajustés sur tous les paramètres du modèle final (25 variables et 6 interactions). Ils ont été calculés pour tous les facteurs présents dans le modèle sauf pour ceux qui avaient été identifiés dans l'analyse précédente (recherche des facteurs d'ajustement) comme modificateur d'effet (pas de sens sur le plan épidémiologique). Il s'agit des variables "Sanitaires", "Indice de niveau économique", "Niveau scolaire de la mère", "Allocations familiales", "Taille de la mère" et "Sexe de l'enfant". Concernant ces dernières il ne sera donc pas possible de faire de comparaison avec les mesures d'association brutes. Lorsqu'un facteur était modificateur d'effet pour un autre, les effectifs et les prévalences ont été désagrégés.

- **Description des résultats**

Les résultats ont été commentés à partir des risques relatifs.

Chez les ménages qui perçoivent des allocations familiales, un enfant a environ 2 fois plus de chances d'avoir un retard de taille si le ménage dispose d'une quantité d'eau insuffisante par rapport à un ménage qui dispose d'eau à volonté. Ce risque n'est plus significatif si le ménage ne perçoit pas d'allocations familiales.

Un enfant dont le chef de ménage réside à Pikine depuis moins de 6 ans a moins de chances d'avoir un retard de taille (RR=0.7) que celui dont le chef de ménage réside à Pikine depuis plus de 13 ans.

Un enfant âgé de 12 à 23 mois et dont la mère mesure moins de 1.60 m a environ 7 fois plus de chances d'avoir un retard de taille que celui âgé de moins de 12 mois. Ce risque de retard de taille est multiplié par 5 si la mère mesure entre 1.60 et 1.65 m. Si la mère mesure plus de 1.65 m, ce risque est multiplié seulement par 3. Par rapport aux enfants de moins de 12 moins, les risques les plus élevés s'observent chez ceux âgés de 12 à 23 mois si la mère mesure moins de 1.60 et ceux âgés de 36 à 47 mois si la mère mesure plus de 1.60.

Une fille âgée de 36 à 47 mois a environ 4 fois plus de chances d'avoir un retard de taille que celle âgée de moins de 12 mois. Chez les garçons ce risque est multiplié par 2 mais n'est plus significatif.

Un enfant dont le ménage ne possède pas de sanitaires privés et qui a eu une diarrhée dans les 15 derniers jours a environ 2 fois plus de chances d'avoir un retard de taille que celui qui n'a pas eu de diarrhée récente. Ce risque n'est plus significatif si le ménage possède des sanitaires privés.

- **Validité des modèles**

La déviance des modèles logistique et log-binomial n'a pu être estimée en raison de faibles effectifs dans le tableau des données regroupées . En effet lorsque le nombre de variables dans le tableau est élevé, le nombre de combinaison de modalités de variables devient tel que les effectifs pour chaque combinaison sont très petits (4452 lignes avec des effectifs de 2 maximum pour notre tableau). On se retrouve presque dans le cas de données individuelles (4477 lignes). On ne peut donc interpréter les tests de validité basés sur la déviance. En revanche on a pu étudier la validité des modèles par le test de Hosmer et Lemeshow. L'hypothèse nulle d'adéquation des modèles n'a pas été rejetée (p=0.64 pour logistique et p=0.57 pour log-binomial). Le détail des tests est fourni à **l'Annexe 13**. Il semblerait donc que nos modèles s'ajustent bien aux données et que l'hypothèse de la distribution binomiale faite sur le retard de taille soit correcte.

- **Convergence du modèle log-binomial**

Les problèmes de convergence du modèle log-binomial n'ont pas été résolus par l'utilisation des paramètres estimés par le modèle logistique. Il a fallu utiliser la méthode proposée par Schouten et al. (1993).

DISCUSSION

- **Recherche des facteurs d'ajustement**

La recherche des facteurs de confusion et des facteurs modificateur d'effet peut se faire indifféremment par l'utilisation du modèle logistique ou du modèle log-binomial. Les résultats obtenus sont les mêmes. Mais il arrive parfois que la fonction de lien **ln** ne permette pas d'estimer immédiatement le maximum de vraisemblance du modèle log-binomial. La solution pour trouver ce maximum est donc de fournir à l'algorithme de recherche des valeurs de départ adéquates pour l'estimation du maximum cette fonction. En pratique ce problème se résout presque toujours mais l'utilisation de l'outil SAS exige des syntaxes longues et répétitives. Faire cette recherche de fonction sans l'avoir automatisée au préalable par une macro peut devenir une tâche contraignante et longue à accomplir. C'est pourquoi à résultats équivalents il est préférable et plus simple d'utiliser le modèle logistique pour la recherche de ces facteurs d'ajustement.

- **Approximation du risque relatif**

Dans la plupart des enquêtes épidémiologiques la quantification de l'association entre un facteur et le risque de maladie se fait par l'utilisation d'odds-ratios en raison de leur caractère "tout terrain" (estimable dans tous les types d'enquêtes). Cette mesure est le plus souvent interprétée comme un risque relatif même dans les cas où il est possible d'estimer des RR. Nous avons comparé tout au long notre étude ces deux indices d'association pour savoir s'il pouvait exister des différences.

Pour la plupart des facteurs de notre étude l'écart entre odds-ratios et risques relatifs est faible même pour des prévalences allant jusqu'à 20% : le risque relatif de retard de taille d'un enfant dont la mère mesure moins de 1.60 m par rapport à celui dont la mère mesure plus de 1.65 m peut être estimé à l'aide d'un odds-ratio (à 19% OR=2.4 et RR=2.1). Mais dans quelques cas cet écart ne devient plus négligeable : chez un enfant âgé entre 36 et 47 mois et dont la mère mesure moins de 1.60 m le risque de retard de taille par rapport à un enfant de moins de 12 mois est sensiblement surestimé par l'OR (à 19% OR=6.9 et RR=5.6). On retrouve cette différence pour toutes les classes d'âge de l'enfant. On retiendra donc que si les prévalences sont faibles (<20%) et si le risque relatif n'est pas trop élevé, il semblerait qu'il soit moins probable de rencontrer des différences importantes entre OR et RR. Mais si ce risque est élevé, il ne faut plus interpréter l'odds-ratio comme une approximation du risque relatif car ce dernier est alors surestimé. Ces constatations avaient également été faites par Zocchetti et al. (1997).

- **Comparaison analyses bi-, tri- et multivariée**

Les indices d'association entre le retard de taille et les facteurs sont en général plus faibles lorsqu'ils sont ajustés simultanément à tous les autres facteurs. Un enfant dont le chef de ménage n'est pas scolarisé a 2.1 fois plus de chances d'avoir un retard de taille que celui dont le chef de ménage possède un brevet et plus; ajusté aux facteurs de confusion, ce risque passe à 1.3 . Il est en de même pour la plupart des autres facteurs excepté l'âge de l'enfant. Par exemple chez les enfants de 12 à 23 mois dont la mère mesure moins de 1.60 m, le risque de retard de taille est multiplié par 3.4, alors qu'ajusté aux facteurs de confusion ce risque passe à 6 9. On observe également cela dans toutes les autres classes d'âges.

De plus la significativité[10] des associations entre le retard de taille et ces facteurs peut varier après ajustement. Un enfant dont le chef de ménage n'est pas scolarisé par rapport à un enfant dont le chef de ménage possède un brevet ou plus a un risque de retard de taille

[10] On se place à l'échelle de la population où on raisonne à partir des intervalles de confiance

Tableau 11 : Intervalles de confiance significatifs à 95% en analyse bivariée et en analyse multivariée

Facteur de risque	Classe d'exposition	Classe de référence	Analyse bivariée	Analyse multivariée
Densité d'individus par pièce	>2.5	<1.5	*	
Source d'eau	Autre	Privée	*	
Quantité d'eau disponible	Insuffisante	A volonté	*	*[1]
Quantité d'eau disponible	Suffisante	A volonté		*[1]
Indice de niveau économique	Faible	Elevé	*	
Age du chef de ménage	50 ans et plus	Moins de 50 ans	*	
Ethnie du chef de ménage	Toucouleur	Wolof	*	
Niveau scolaire du chef de ménage	Non scolarisé	Brevet et plus	*	
	Avant brevet	Brevet et plus	*	
Taille de la mère	Moins des 1.60 m	Plus de 1.65 m	*	
	Entre 1.60 et 1.65 m	Plus de 1.65 m	*	
Ethnie de la mère	Autre	Wolof	*	
Niveau scolaire de la mère	Non scolarisée	Scolarisée	*	
Parenté de la mère avec le chef de ménage	Autre	Conjoint	*	
Enfants hospitalisés pour malnutrition grave	Oui	Non	*	
Allocations familiales	Non	Oui	*	
Age de l'enfant	De 12 à 23 mois	Moins de 12 mois	*	*[2]
	De 24 à 35 mois	Moins de 12 mois	*	*[2]
	De 36 à 47 mois	Moins de 12 mois	*	*[2]
	De 48 à 59 mois	Moins de 12 mois	*	*[2]
Poids de naissance	Inconnu	Connu	*	
Parenté de l'enfant avec le chef de ménage	Autre	Fils/ Fille	*	
Dernière diarrhée récente	Oui	Non	*	*

[1] Dans la strate Allocations familiales perçues

[2] Dans les trois strates de la variable "Taille de la mère" et dans la strate "Fille" de la variable "sexe de l'enfant"

significatif en analyse brute ($IC_{95\%}$=[1.54-2.88]). Ce risque n'est plus significatif lorsqu'il est ajusté sur les facteurs de confusion ($IC_{95\%}$=[0.93-1.86]). Cette significativité varie d'autant plus que les indices sont proches de 1. Le **Tableau 11** compare les facteurs de risque significatifs en analyse bivariée et multivariée. Pour chaque facteur la classe d'exposition ainsi que celle de référence ont été précisées. L'analyse de ce tableau montre que les facteurs de risque retenus dans la première analyse ne sont plus significatifs pour la plupart dans la deuxième. Cette constatation souligne particulièrement le caractère indispensable de ces deux analyses complémentaires. En effet si l'analyse brute permet d'identifier les facteurs liés au risque de retard de taille, l'analyse multivariée permet de savoir si ces facteurs sont toujours associés à cet état de santé et donc si l'effet de ces facteurs sur le risque de retard de taille est intrinsèque ou alors dû à un ensemble de facteurs. L'exemple ci-dessus illustre bien ce propos : le risque de retard de taille de l'enfant est-il dû au fait que le chef de ménage ne soit pas scolarisé ou alors au contexte lié à cet état ?

- **Confusion résiduelle**

Le choix des variables à ajouter aux facteurs de risque significatifs dans les modèles multivariés ne s'est pas limité aux seuls facteurs de confusion et modificateur d'effet. On a volontairement gardé les autres facteurs afin de prendre en compte tout effet de confusion résiduel. Pour pouvoir juger de l'importance de cette confusion, nous avons comparé les modèles finaux (25 variables et 6 interactions) avec des modèles de régression trivariés incluant chaque facteur de risque et soit son facteur de confusion soit son facteur modificateur d'effet. Les résultats sont présentés aux **Annexes 8** et **9**.

Un enfant dont le chef de ménage est âgé de 50 ans et plus a 1.2 fois plus de chances d'avoir un retard de taille que celui dont le chef de ménage est âgé de moins de 50 ans. Ajusté au niveau scolaire de ce chef de ménage, ce risque passe à 1.1 et n'est plus significatif. Ajusté aussi sur les facteurs de confusion résiduels ce risque reste non significatif. Cette situation est rencontrée pour la plupart des facteurs de risque de notre étude.

Un enfant âgé de 24 à 35 mois et dont la mère mesure plus de 1.65 m a environ 2 fois plus de chances d'avoir un retard de taille que celui âgé de moins de 12 mois. Ajusté sur les facteurs de confusion résiduels ce risque est multiplié par 4.

Une fille âgée entre 12 et 23 mois a environ 5 fois plus de chances d'avoir un retard de taille que celle âgée de moins de 12 mois. Ajusté aux facteurs de confusion résiduels ce risque passe à 3. Ces deux dernières constatations ont été faites pour toutes les classes d'âge de l'enfant et quel que soit son sexe.

Les facteurs de confusion résiduels modifient peu ou pas la plupart des associations risque retard de taille / facteur de risque. En revanche pour certains facteurs la non-prise en compte de cette confusion peut masquer sensiblement la réalité de l'association. La conséquence sur les résultats se traduit par des indices trop élevés ou pas assez forts par rapport à la réalité de l'échantillon.

CONCLUSION

L'analyse brute de l'association du retard de taille avec les descripteurs socio-économiques, sanitaires et familial dans lequel vit l'enfant a montré qu'un grand nombre de facteurs était lié à cet état de santé. Ce résultat, en accord avec le schéma général décrit à la **Figure 1**, confirme bien le caractère multifactoriel du retard de taille.

La recherche des facteurs d'ajustement a nécessité l'utilisation de macros SAS. De nombreux facteurs de confusion et modificateur d'effet ont été ainsi mis en évidence. L'ajustement sur ces facteurs a sensiblement modifié la nature des associations précédemment décrites. De plus la comparaison de modèles de régression tri- et multivariés a permis de souligner l'importance de la prise en compte de la confusion résiduelle dans l'étude de ces associations.

Le test de Hosmer et Lemeshow a montré une bonne adéquation des modèles aux données.

La comparaison des modèles de régressions logistique et log-binomial a permis de mettre en évidence quelques différences importantes entre les mesures d'association dans les cas où ces mesures étaient élevées. L'utilisation de l'odds-ratio comme une approximation du risque relatif peut surestimer sensiblement la valeur du risque relatif. Lorsque le type d'étude le permet (cohorte et transversale), il est donc préférable d'utiliser des modèles de régression log-binomial qui permettent d'estimer directement des risques relatifs.

Cette étude montre aussi que d'autres questions importantes nécessitent d'être traitées.

Il s'agit tout d'abord du risque global de première espèce dans la recherche des facteurs d'ajustement (analyse trivariée). Lorsqu'on cherche à identifier des facteurs de confusion à partir des données de l'étude, le nombre important de tests effectués fait que l'on ne sait plus interpréter le risque α. En effet faire beaucoup de tests augmente la probabilité de trouver des "faux positifs", c'est-à-dire la probabilité de trouver des associations significatives par hasard. Une solution possible est de choisir un seuil α très petit, par exemple α/n, si n tests sont effectués. Procéder ainsi a tout de même l'inconvénient d'augmenter le nombre de "faux négatifs", c'est-à-dire la probabilité de passer à coté de certains facteurs de risque.

Une autre question est la prise en compte de l'effet du plan de sondage dans les modèles. Le risque relatif et l'odds-ratio ont été définis comme des mesures de l'association entre un facteur de risque et la maladie. Il s'agit d'indices définis au niveau de la population. Leur calcul exact nécessiterait de connaître la répartition de tous les sujets de la population selon le facteur de risque et la maladie. En pratique cette information a été recueillie sur un échantillon représentatif de cette population obtenu par un sondage en grappes à deux degrés après stratification. Ne disposant pas d'outil performant, nous avons traité ces données comme si elles provenaient d'un sondage aléatoire simple et cela malgré le biais possible sur la précision des valeurs obtenues. Or un biais sur la précision de ces résultats peut entraîner de fausses conclusions sur la significativité de certains facteurs de risque. Une étude possible que l'on pourrait effectuer dans le prolongement de ce travail serait l'évaluation de l'effet du plan de sondage en grappes sur la précision des estimations des indices d'association obtenue à partir des modèles de régressions logistique et log-binomial.

Il existe enfin d'autres méthodes possibles pour étudier une variable réponse de type dichotomique : l'analyse discriminante et l'arbre de discrimination type CART. La première, basée sur une analyse factorielle et conçue initialement pour des variables quantitatives, peut être utilisée pour l'étude de descripteurs qualitatifs moyennant des étapes de codages préliminaires. Toutefois l'interprétation finale n'est pas facile. La seconde, basée sur la construction d'arbres de décision binaires, a l'avantage de la lisibilité immédiate des résultats. Elle permet également de prendre en compte les interactions de manière implicite. Ces deux méthodes ont le désavantage commun de ne pas permettre une interprétation des résultats en termes d'indices d'association épidémiologiques. Toutefois l'utilisation conjointe de ces trois

techniques de discrimination pourraient améliorer l'identification des facteurs liés au risque de retard de taille.

BIBLIOGRAPHIE

Bouyer J. (1991). La régression logistique en Epidémiologie (parties 1 et 2). *Rev. Epidemiol. Santé Publ.*, 39, 79-87, 183-196.

Bouyer J., Hémon D., Cordier S. et al. (1995). *Epidémiologie. Principes et méthodes quantitatives*. Editions INSERM, Paris. 272 pp.

Dagnelie P. (1973). *Théories et Méthodes Statistiques*, Volume 1. Les Presses Agronomiques de Gembloux. 377 pp.

Dean A.G., Dean J.A., Coulombier D. et al. (1994). Epi-Info Version 6, *A Word processing, Database and Statistics program for Public Health on IBM compatible micros-computers*. Atlanta, Georgia : Center for Disease Control and Prevention. 601 pp.

Dobson J.A. (1990). *An introduction to Generalized Linear Model*. Chapman & Hall, London. 174 pp.

Hosmer D.W. et Lemeshow S. (1989). *Applied Logistic Regression*. A Wiley series in Prob. And Math. Stat., New York. 307 pp.

OMS (1983). *Mesure des modifications de l'état nutritionnel*. Genève : Organisation mondiale de la Santé. 104 pp.

OMS (1995). Utilisation et interprétation de l'anthropométrie. *Série de rapports techniques*, 854. Genève : Organisation Mondiale de la Santé. 449 pp.

Rothman K.J. (1986). *Modern Epidemiology*. Boston, Little Brown & co. 358 pp.

SAS Institute Inc. (1989a). *SAS Language and Procedures : Usage*, Version 6, First Edition. Cary, NC : SAS Institute Inc. 638 pp.

SAS Institute Inc. (1989b). *SAS/STAT User's guide*, Version 6, Fourth Edition, Volumes 1 & 2. Cary, NC : SAS Institute Inc. 943 pp. & 846 pp.

SAS Institute Inc. (1990a). *SAS Guide to Macros Processing*, Version 6, Second Edition. Cary, NC : SAS Institute Inc. 319 pp.

SAS Institute Inc. (1990b). *SAS Procedures Guide*, Version 6, Third Edition. Cary, NC : SAS Institute Inc. 705 pp.

SAS Institute Inc. (1995). *Logistic Regression Examples Using the SAS System*, Version 6, First Edition. Cary, NC : SAS Institute Inc.163 pp.

SAS Institute Inc. (1997). *SAS/STAT Software : changes and enhancements through Release 6.12*. Cary N.C.: SAS Institute Inc. 1167 pp.

Schouten E.G., Dekker J.M., Kok F.J. et al. (1993). Risk ratio and rate ratio estimation in case-cohort design : hypertension and cardiovascular mortality. *Stat. Med.*,12, 1743-45.

Simonoff J. S. (1998). Logistic Regression, Categorical Predictors, and Goodness of it : It depends on who you ask. *The American Stat.*, vol. 52, n°1.

Traissac P., Delpeuch F., Maire B. (1997). Construction d'un indice synthétique de niveau économique des ménages dans les enquêtes nutritionnelles. Exemples d'application au Congo. *22ième Congrés des Epidémiologistes de Langue Française*, Montpellier, 2-4 avril 1997. Résumé publié dans *Rev. Epidemiol. Santé Publ.*, 45, S114-S115.

Unité de Nutrition de l'Orstom (1997). Résultats de l'enquête nutritionnelle menée à Pikine en mai-juin 1996. *Rapport d'étude*. Dakar, 51 pp.

Zocchetti C., Consonni D., Bertazzi P.A. (1997). Relationship between prevalence rate ratios and odds-ratios in cross-sectionnal studies. *Internat. J. Epidemiol.*, 26(1), 220-223.

ANNEXES

Annexe 1 : Construction de l'indice de niveau économique : résultats de l'AFC

7a Inertie des axes factoriels

0.50880	0.25888	9869.17	17.26%	**********
0.40711	0.16574	6318.49	11.05%	**********
0.34075	0.11611	4426.61	7.74%	********
0.32502	0.10564	4027.19	7.04%	*******
0.31212	0.09742	3713.91	6.49%	*****
0.30204	0.09123	3478.02	6.08%	****
0.28874	0.08337	3178.46	5.56%	***
0.28364	0.08045	3067.01	5.36%	***
0.27872	0.07769	2961.65	5.18%	****
0.26566	0.07058	2690.63	4.71%	***
0.25784	0.06648	2534.40	4.43%	***
0.24180	0.05847	2229.04	3.90%	***
0.22652	0.05131	1956.20	3.42%	****
0.21387	0.04574	1743.73	3.05%	***
0.19665	0.03867	1474.25	2.58%	***
0.19324	0.03734	1423.55	2.49%	**
0.18138	0.03290	1254.22	2.19%	**
0.14829	0.02199	838.28	1.47%	*
1.50000		57184.8 (Degrees of Freedom 841)		

7b Contributions partielles à l'inertie et cosinus carrés des variables

		Contributions absolues (CTA*1000)			Cosinus carrés (CTR*1000)		
Variables	**Modalités**	**Axe 1**	**Axe 2**	**Axe 3**	**Axe 1**	**Axe 2**	**Axe 3**
Qualité du logement	propre	1	2	69	7	14	336
	sale	53	9	204	182	20	317
	tres_propre	48	26	33	184	63	56
Matériau des murs	mur_autre	20	19	148	65	40	213
	mur_brique	1	1	4	65	40	213
Matériau du toit	toit_autre	24	92	3	135	337	8
	toit_tuile_ciment	20	78	3	135	337	8
Matériau du sol	sol_ciment	0	0	10	0	0	34
	sol_terre	21	108	44	80	263	76
	sol_carrele	19	86	1	77	225	1
Nombres de pièces	moins_de_3	59	53	19	265	152	38
	entre_3_et_6	0	3	7	0	10	16
	plus_de_6	55	25	2	263	77	4
Source d'eau	eau_privee	74	0	7	326	1	13
	eau_autre	31	0	3	326	1	13
Sanitaires	wc_autre	40	5	62	222	18	154
	wc_prive	31	4	48	222	18	154
Electricité	elec_oui	101	7	2	462	20	3
	elec_non	48	3	1	462	20	3
Combustible domestique	bois_charbon	50	3	66	228	10	136
	gaz_petrole	23	2	31	228	10	136
Indice de biens	biens_1	86	0	59	339	0	105
	biens_2	18	1	128	70	1	220
	biens_3	0	2	9	1	4	15
	biens_4	26	1	4	102	1	6
	biens_5	63	1	17	246	3	29
Possession de mouton avant tabaski	moutav_non	11	98	5	96	572	19
	moutav_oui	20	189	9	96	572	19
Possession de mouton après tabaski	moutap_non	12	38	0	178	364	1
	moutap_oui	46	145	1	178	364	1

7c Représentation des variables dans le plan factoriel (1,2)

Annexe 2 : Tests de liaison des 38 facteurs avec le retard de taille

	Tests du khi² n=4567 avec toutes les u.s.[11]
Caractéristiques du ménage	
Taille du ménage	P=0.71
Nombre d'enfants de moins de 5 ans	P=0.85
Densité d'individus par pièce	**P<0.01**
Statut d'occupant du ménage	P=0.50
Source d'eau	**P<0.01**
Quantité d'eau disponible	**P=0.01**
Sanitaires	**P=0.13**
Indice de niveau économique	**P<0.01**
Caractéristiques du chef de ménage	
Sexe	**P=0.17**
Age	**P=0.02**
Ethnie	**P=0.15**
Niveau scolaire	**P<0.01**
Secteur d'activité	**P<0.01**
Durée de résidence à Pikine	**P=0.01**
Caractéristiques de la mère	
Mère biologique	P=0.94
Taille	**P<0.01**
Age	P=0.88
Ethnie	**P=0.12**
Niveau scolaire	**P<0.01**
Nombre d'enfants de moins de 5 ans	P=0.58
Situation matrimoniale	**P=0.11**
Occupation de la mère	P=0.54
Parenté avec le chef de ménage	**P=0.06**
Enfants hospitalisés pour malnutrition grave	**P<0.01**
Allocations familiales	**P=0.03**
Caractéristiques de l'enfant	
Sexe	**P=0.16**
Age	**P<0.01**
Connaissance du poids à la naissance	**P<0.01**
Rang de l'enfant dans la fratrie	P=0.91
Parenté avec le chef de ménage	**P=0.03**
Aliments interdits	P=0.51
Dernière diarrhée récente	**P<0.01**
Dernière diarrhée traitée avec médicaments	P=0.24
Dernière diarrhée traitée avec pain de singe	**P=0.09**
Dernière diarrhée traitée avec la solution sucrée/salée	P=0.57
Dernière diarrhée traitée avec sachet unicef	P=0.42
Existence d'un carnet de vaccination	**P=0.14**
Existence d'un carnet de suivi de croissance	P=0.59

[11] Unités statistiques

Annexe 3 : Comparaison de distributions entre le groupe d'enfants sans données manquantes (n=4477) et le groupe d'enfants exclus de l'analyse multivariée (n=90)

		Tests du Khi²
Variable expliquée	Retard de taille	P=0.30
Caractéristiques du ménage	Densité d'individus par pièce	P=0.84
	Source d'eau	P=0.23
	Quantité d'eau disponible	**P=0.02**
	Sanitaires	**P<0.01**
	Indice de niveau économique	P=0.36
Caractéristiques du chef de ménage	Sexe	P=0.33
	Age	P=0.07
	Ethnie	P=0.11
	Niveau scolaire	P=0.09
	Secteur d'activité	**P=0.04**
	Durée de résidence à Pikine	P=0.23
Caractéristiques de la mère	Taille	P=0.69
	Ethnie	**P<0.01**
	Niveau scolaire	P=0.65
	Situation matrimoniale	P=0.20
	Parenté avec le chef de ménage	**P=0.02**
	Enfants hospitalisés pour malnutrition grave	P=0.31
	Allocations familiales	**P=0.03**
Caractéristiques de l'enfant	Sexe	P=0.59
	Age	P=0.11
	Connaissance du poids de naissance	P=0.77
	Parenté avec le chef de ménage	**P<0.01**
	Dernière diarrhée récente	P=0.55
	Dernière diarrhée traitée avec pain de singe	P=0.14
	Existence d'un carnet de vaccination	P=0.35

Annexe 4 : Indices d'association bruts du retard de taille avec les facteurs de risque potentiels

4a. Caractéristiques du ménage

Facteur	n		Prévalence %	Odds-ratio brut ($IC_{95\%}$)	Risque relatif brut ($IC_{95\%}$)
Densité d'individus par pièce					
< 1.5	960	21%	10.1	1	1
Entre 1.5 et 2.5	1838	41%	12.4	1.26 (0 98-1 62)	1.23 (0 98-1 54)
> 2.5	1679	38%	15.2	1.60 (1 24-2 05)	1.51 (1 21-1 88)
Source d'eau					
Privée	3174	71%	11.7	1	1
Autre	1303	29%	16.2	1.46 (1 22-1 76)	1.39 (1 19-1 62)
Quantité d'eau disponible					
A volonté	1553	35%	11.3	1	1
Suffisante	1773	40%	13.0	1.17 (0.95-1 45)	1.15 (0 96-1 38)
Insuffisante	1151	26%	15.3	1.42 (1 13-1 78)	1.36 (1.12-1 65)
Sanitaires					
Privés	2720	61%	12.3	1	1
Autres	1757	39%	14.0	1.16 (0 97-1 38)	1.14 (0 97-1.32)
Indice de niveau économique (terciles)					
Elevé	1601	36%	10.2	1	1
Moyen	1511	34%	12.1	1.22 (0.97-1 52)	1.19 (0 97-1 45)
Faible	1365	30%	17.2	1.83 (1 48-2 27)	1.69 (1 40-2 04)

4b. Caractéristiques du chef de ménage

Facteur	n		Prévalence %	Odds-ratio brut ($IC_{95\%}$)	Risque relatif brut ($IC_{95\%}$)
Sexe					
Homme	3976	89%	12.7	1	1
Femme	501	11%	14.8	1.19 (0 91-1 54)	1.16 (0 92-1 45)
Age					
Moins de 50 ans	2572	57%	12.0	1	1
50 ans et plus	1905	43%	14.3	1.23 (1 03-1 46)	1.20 (1 03-1 39)
Ethnie					
Wolof	1879	42%	11.7	1	1
Toucouleur	757	17%	14.5	1.29 (1 01-1 65)	1.25 (1.01-1.54)
Serer	564	13%	13.2	1.14 (0 86-1 52)	1.13 (0 88-1 44)
Autre	1277	29%	14.0	1.23 (0 99-1 52)	1.20 (0 99-1 44)
Niveau scolaire					
Non scolarisé	2899	65%	14.8	2.30 (1 66-3 23)	2.10 (1 54-2 88)
Avant brevet	1022	23%	11 1	1.66 (1 14-2 43)	1.59 (1 12-2 25)
Brevet et plus	556	12%	7.0	1	1
Secteur d'activité					
Etat	813	18%	9.1	0 67 (0 50-0 89)	0.70 (0 54-0 90)
Privé	1473	33%	13 0	1	1
Informel	1436	32%	13 7	1.06 (0 86-1 31)	1.05 (0 87-1 27)
Agriculture/autre	755	17%	15.6	1 24 (0 96-1 58)	1.20 (0 97-1 48)
Durée de résidence à Pikine					
Moins de 6 ans	803	18%	9 7	0.68 (0 53-0 88)	0.71 (0 57-0 90)
Entre 6 et 13 ans	925	21%	14 0	1.03 (0 83-1 28)	1.03 (0 85-1 23)
Plus de 13 ans	2749	61%	13 6	1	1

4c. Caractéristiques de la mère

Facteur	n		Prévalence %	Odds-ratio brut ($IC_{95\%}$)	Risque relatif brut ($IC_{95\%}$)
Taille					
Moins de 1.60 m	1313	29%	18.6	2.41 (1 94-3 00)	2.15 (1 78-2 60)
Entre 1.60 et 1.65	1393	31%	13.2	1.61 (1.28-2.02)	1.53 (1 25-1 87)
Plus de 1.65 m	1771	40%	8.6	1	1
Ethnie					
Wolof	1879	42%	11.7	1	1
Toucouleur	741	17%	13.4	1.17 (0 91-1 51)	1.15 (0 92-1 43)
Serer	564	13%	14.4	1.27 (0 97-1 67)	1.23 (0 97-1 56)
Autre	1293	29%	14.1	1.24 (1.01-1 53)	1.21 (1 01-1 45)
Niveau scolaire					
Scolarisée	1340	30%	9.7	1	1
Non scolarisée	3137	70%	14.4	1.56 (1 27-1 92)	1.48 (1 23-1 78)
Situation matrimoniale					
Mariée	4057	91%	12.7	1	1
Autre	420	9%	15.5	1.26 (0 95-1.66)	1.22 (0.96-1.54)
Parenté avec le chef de ménage					
Chef de ménage	120	3%	15.8	1.38 (0.83-2.28)	1.32 (0 86-2 02)
Conjoint	2715	61%	12.0	1	1
Autre	1642	37%	14.4	1.23 (1.03-1.47)	1.20 (1 02-1.40)
Enfants hospitalisés pour malnutrition grave					
Oui	628	14%	16.7	1.42 (1.13-1 79)	1.35 (1 11-1 64)
Non	3849	86%	12.4	1	1
Allocations familiales					
Oui	586	13%	10.1	1	1
Non	3891	87%	13.4	1.38 (1 04-1 84)	1.33 (1 03-1 72)

4d. Caractéristiques de l'enfant

Facteur	n		Prévalence %	Odds-ratio brut ($IC_{95\%}$)	Risque relatif brut ($IC_{95\%}$)
Sexe					
Fille	2217	50%	12.3	1	1
Garçon	2260	50%	13.6	1.12 (0 94-1 34)	1.11 (0 95-1.29)
Age					
Moins de 12 mois	883	20%	6.1	1	1
De 12 à 23 mois	938	21%	15.6	2.83 (2 04-3 92)	2.55 (1 89-3 43)
De 24 à 35 mois	933	21%	13.2	2.33 (1 67-3.26)	2.16 (1 59-2.93)
De 36 à 47 mois	879	20%	17.0	3.13 (2 26-4 34)	2.77 (2 06-3 73)
De 48 à 59 mois	844	19%	12.9	2.28 (1 62-3 20)	2.11 (1 54-2.88)
Poids de naissance					
Connu	2902	65%	11.9	1	1
Inconnu	1575	35%	15 0	1.31 (1 09-1 56)	1.26 (1 08-1 47)
Parenté avec le chef de ménage					
Fils/fille	2750	61%	12.1	1	1
Petit(e) fils/fille	1156	26%	13 7	1.16 (0 94-1 42)	1.14 (0 95-1 35)
Autre	571	13%	15.6	1.34 (1 04-1 73)	1 29 (1 04-1 60)
Dernière diarrhée récente					
Oui	1221	27%	16.0	1 41 (1 17-1 70)	1.35 (1 15-1 58)
Non	3256	73%	11 9	1	1
Diarrhée traitée avec pain de singe					
Oui	1970	44%	14.1	1	1
Non	2507	56%	12 1	0.84 (0 70-1 00)	0.86 (0 73-1 00)
Carnet de vaccination					
Oui	3561	80%	12 6	1	1
Non	916	20%	14 5	1.18 (0 96-1 45)	1.15 (0 96-1 38)

Annexe 5 : Liste des facteurs modificateurs d'effet obtenue en analyse trivariée (α=5%)[12]

Facteur de risque	Facteurs modificateurs d'effet	p-value du terme d'interaction	
		logit	log
Age de l'enfant	Sexe de l'enfant	0.0386	0.0349
	Taille de la mère	0.0183	0.0225
Connaissance du poids de naissance	Parenté de l'enfant avec le chef de ménage	0.0109	0.0124
	Parenté de la mère avec le chef de ménage	0.0185	0.0210
Densité d'individus par pièce	Indice de niveau économique	0.0113	0.0107
Dernière diarrhée récente	Densité d'individus par pièce	0.0364	0.0389
	Sanitaires	0.0036	0.0041
Quantité d'eau disponible	Allocations familiales	0.0149	0.0137
Indice de niveau économique	Secteur d'activité du chef de ménage	0.0303	0.0308
Secteur d'activité du chef de ménage	Indice de niveau économique	0.0303	0.0308
Sanitaires	Niveau scolaire de la mère	0.0147	0.0137
Source d'eau	Niveau scolaire de la mère	0.0134	0.0110

Annexe 6 : Liste des facteurs de confusion obtenue en analyse trivariée (α=5%)[13]

Facteur de risque	Facteur de confusion	p-value du terme de confusion			
		logit		log	
		Test 1*	Test 2**	Test 1	Test 2
Age du chef de ménage	Durée de résidence du chef de ménage à Pikine	0.0208	0.0639	0.0208	0.0644
	Niveau scolaire de chef de ménage	0.0208	0.2766	0.0208	0.2795
	Parenté de la mère avec le chef de ménage	0.0208	0.0884	0.0208	0.0915
Allocations familiales	Indice de niveau économique	0.0206	0.1338	0.0206	0.1452
	Niveau scolaire de chef de ménage	0.0206	0.2442	0.0206	0.2453
	Niveau scolaire de la mère	0 0206	0.0663	0.0206	0.0688
	Secteur d'activité du chef de ménage	0.0206	0 1464	0.0206	0.1519
Quantité d'eau disponible	Indice de niveau économique	0 0092	0 6190	0.0092	0.6554
	Source d'eau	0 0092	0.4288	0.0092	0.4336
Secteur d'activité du chef de ménage	Niveau scolaire de chef de ménage	0.0007	0.1982	0 0007	0 2010
Source d'eau	Indice de niveau économique	0.0001	0.2353	0.0001	0.2613

[12] Le détail des indices d'association ajustés sur ces facteurs modificateur d'effet est donné à **l'Annexe 8.**
[13] Le détail des indices d'association ajustés sur ces facteurs de confusion est donné à **l'Annexe 9**
* P-value extraite de la rubrique "LR Statistics for Type1"
** P-value extraite de la rubrique "LR Statistics for Type3"

Annexe 7 : Indices d'association du retard de taille ajustés (analyse multivariée)

7a. Caractéristiques du ménage

Facteur modificateur d'effet		Facteur de risque	n		Prévalence %	Odds-ratio ajusté ($IC_{95\%}$)	Risque relatif ajusté ($IC_{95\%}$)
		Densité d'individus par pièce					
Indice de niveau économique	*Faible*	**< 1.5**	164	4%	18 9	1	1
		Entre 1.5 et 2.5	407	9%	13 5	0 64 (0 26-1 60)	0 69 (0 32-1 50)
		> 2.5	794	18%	18 8	0 93 (0 34-2.55)	0 93 (0 39-2 21)
	Moyen	**< 1.5**	283	6%	10 2	1	1
		Entre 1.5 et 2.5	648	14%	11.6	1 18 (0 48-2 89)	1 11 (0 51-2 43)
		> 2.5	580	13%	13 6	1 52 (0 55-4 17)	1 38 (0 57-3 35)
	Elevé	**< 1.5**	513	11%	7 2	1	1
		Entre 1.5 et 2.5	783	17%	12 5	1 63 (1 08-2 46)	1 53 (1 07-2 20)
		> 2.5	305	7%	9 2	1 16 (0 68-1 97)	1 16 (0 73-1 86)
		Source d'eau					
--		**Privée**	3174	71%	11 7	1	1
		Autre	1303	29%	16 2	1 32 (0 94-1 86)	1 22 (0 93-1 59)
		Quantité d'eau disponible					
Allocations familiales	*Oui*	**A volonté**	253	6%	5.5	1	1
		Suffisante	227	5%	11 4	2.27 (1 13-4 58)	1 99 (1 08-3 67)
		Insuffisante	106	2%	17 9	2 45 (1 08-5 57)	2.28 (1 16-4 52)
	Non	**A volonté**	1300	29%	12 4	1	1
		Suffisante	1546	35%	13 2	0 96 (0 27-3 47)	0 96 (0 31-2 99)
		Insuffisante	1045	23%	15 0	0 83 (0 20-3.34)	0 88 (0.27-2 91)

7b. Caractéristiques du chef de ménage

Facteur modificateur d'effet	Facteur de risque	n		Prévalence %	Odds-ratio ajusté ($IC_{95\%}$)	Risque relatif ajusté ($IC_{95\%}$)
	Sexe					
--	**Homme**	3976	89%	12 7	1	1
	Femme	501	11%	14 8	0 96 (0 67-1 37)	1 02 (0 77-1 36)
	Age					
--	**Moins de 50 ans**	1905	43%	14 3	1	1
	50 ans et plus	2572	57%	12 0	1 14 (0 91-1 42)	1.13 (0 94-1 35)
	Ethnie					
--	**Wolof**	1879	42%	11 7	1	1
	Toucouleur	757	17%	14 5	1 41 (0 96-2 06)	1 33 (0 97-1 82)
	Serer	564	13%	13 2	0 94 (0 63-1 39)	0 92 (0 67-1 28)
	Autre	1277	29%	14 0	1 13 (0 81-1 57)	1 15 (0 88-1 51)
	Secteur d'activité					
	Etat	813	18%	9 1	0 75 (0 55-1 03)	0 80 (0 61-1 04)
	Privé	1473	33%	13 0	1	1
	Informel	1436	32%	13 7	0 94 (0 74-1 20)	0 94 (0 77-1 14)
	Agriculture/ autre	755	17%	15 6	1 05 (0 79-1 39)	1 06 (0 86-1 33)
	Niveau scolaire					
--	**Non scolarisé**	2899	65%	14 8	1 36 (0 91-2 03)	1 31 (0 93-1 86)
	Avant brevet	1022	23%	11 1	1 19 (0 80-1 79)	1 18 (0 82-1 68)
	Brevet et plus	556	12%	7 0	1	1
	Durée de résidence à Pikine					
--	**Moins de 6 ans**	803	18%	9 7	0 72 (0 54-0 95)	0 75 (0 60-0 96)
	Entre 6 et 13 ans	925	21%	14 0	1 03 (0 81-1 30)	1 04 (0 86-1 26)
	Plus de 13 ans	2749	61%	13 6	1	1

7c. Caractéristiques de la mère

Facteur modificateur d'effet	Facteur de risque	n		Prévalence %	Odds-ratio ajusté ($IC_{95\%}$)	Risque relatif ajusté ($IC_{95\%}$)
	Ethnie					
	Wolof	1879	42%	11 7	1	1
	Toucouleur	741	17%	13 4	0 89 (0 60-1 32)	0.93 (0 67-1 29)
--	**Serer**	564	13%	14 4	1 18 (0 80-1 73)	1 14 (0 84-1 56)
	Autre	1293	29%	14 1	1 02 (0 74-1 42)	0 98 (0 75-1 29)
	Situation matrimoniale					
	Mariée	4057	91%	12 7	1	1
--	**Autre**	420	9%	15 5	1 09 (0 79-1 51)	1 11 (0 85-1 43)
	Parenté avec le chef de ménage					
--	**Chef de ménage**	120	3%	15 8	1 40 (0 72-2 70)	1.20 (0 72-2 02)
	Conjoint	2715	61%	12 0	1	1
	Autre	1642	37%	14 4	1 38 (0 86-2 20)	1 33 (0 92-1 91)
	Enfants hospitalisés pour Malnutrition grave					
	Oui	628	14%	16 7	1 29 (1 01-1 64)	1 20 (0 99-1 45)
--	**Non**	3849	86%	12.4	1	1

7d. Caractéristiques de l'enfant

Facteur modificateur d'effet		Facteur de risque	n		Prévalence %	Odds-ratio ajusté ($IC_{95\%}$)	Risque relatif ajusté ($IC_{95\%}$)
		Age					
		De 0 à 11 mois	271	6%	7 0	1	1
	Moins de 1 60 m	**De 12 à 23 mois**	273	6%	24 2	9 02 (2 35-34 67)	6 91 (2.03-23 50)
		De 24 à 35 mois	273	6%	19.0	6 95 (1 82-26 45)	5 63 (1 67-18 98)
		De 36 à 47 mois	253	6%	20 9	6 82 (1 80-25 87)	5 31 (1 59-17 78)
		De 48 à 59 mois	243	5%	22 2	7 69 (1 97-29 95)	5 98 (1 74-20 51)
		De 0 à 11 mois	273	6%	6 6	1	1
Taille de la mère	*Entre 1 60 et 1 65 m*	**De 12 à 23 mois**	296	7%	16 9	5 78 (1 47-22 72)	4 89 (1 41-17 03)
		De 24 à 35 mois	288	6%	12 1	4 22 (1 08-16 54)	3 72 (1 07-12 93)
		De 36 à 47 mois	264	6%	21 6	7 50 (1 95-28 84)	5 88 (1 73-19 96)
		De 48 à 59 mois	272	6%	8 9	2 76 (0 68-11 26)	2 59 (0 72-9 37)
		De 0 à 11 mois	339	8%	5 0	1	1
	Plus de 1 65 m	**De 12 à 23 mois**	369	8%	8 1	3 14 (1 40-7 04)	2 98 (1 41-6 27
		De 24 à 35 mois	372	8%	9 7	4 06 (1 84-8 97)	3.61 (1 74-7 52)
		De 36 à 47 mois	362	8%	10.8	4 06 (1 84-8 98)	3.68 (1 77-7 66)
		De 48 à 59 mois	329	7%	9 4	3 64 (1 61-8 23)	3.35 (1 58-7 13)
		Age					
		De 0 à 11 mois	399	9%	3 3	1	1
		De 12 à 23 mois	474	11%	16 0	3 14 (1 40-7 04)	2 98 (1 41-6 27)
	Fille	**De 24 à 35 mois**	468	10%	13 2	4 06 (1 84-8 97)	3 62 (1 74-7 52)
		De 36 à 47 mois	448	10%	15 6	4 06 (1 84-8 98)	3 68 (1 77-7 66)
Sexe de l'enfant		**De 48 à 59 mois**	428	10%	12 1	3 64 (1 61-8 23)	3 35 (1 58-7 13)
		De 0 à 11 mois	484	11%	8 5	1	1
		De 12 à 23 mois	464	10%	15 0	1 06 (0 27-4 07)	1 05 (0 30-3 66)
	Garçon	**De 24 à 35 mois**	465	10%	13 1	1 37 (0 36-5 25)	1 31 (0 37-4 58)
		De 36 à 47 mois	431	10%	18 3	1 76 (0 46-6 68)	1 67 (0 48-5 79)
		De 48 à 59 mois	416	9%	13 7	1 53 (0 39-6 02)	1.47 (0 41-5 23)
		Parenté avec le chef de ménage					
		Fils/fille	2750	61%	12.1	1	1
–		Petit(e) fils/fille	1156	26%	13.7	1 00 (0 61-1 64)	0 80 (0 55-1 17)
		Autre	571	13%	15.6	0 82 (0 51-1 33)	0 97 (0 66-1 43)
		Poids de naissance					
–		**Connu**	2902	65%	11 9	1	1
		Inconnu	1575	35%	15 0	1 25 (0 99-1 57)	1 19 (0 99-1 43)
		Dernière diarrhée récente					
	Privés	**Oui**	705		13 0	1 06 (0 80-1 39)	1 06 (0 85-1 32)
Sanitaires		**Non**	2015		12 0	1	1
	Autres	**Oui**	516		20 0	2 01 (1 20-3 35)	1 74 (1 16-2 62)
		Non	1241		11 5	1	1
		Diarrhée traité avec pain de singe					
–		**Oui**	1970	44%	14 1	1	1
		Non	2507	56%	12 1	0 92 (0 76-1 11)	0 92 (0 79-1 07)
		Carnet de vaccination					
–		**Oui**	3561	80%	12 6	1	1
		Non	916	20%	14 5	0 86 (0 65-1 13)	0 89 (0 71-1 10)

Annexe 8 : Comparaison de l'ajustement par les facteurs modificateurs d'effet en analyse tri- et multivariée

8a. Comparaison des Odds-ratios ajustés

Facteur modificateur d'effet		Facteur de risque	n		Prévalence %	Odds-ratio ajusté ($IC_{95\%}$) Analyse trivariée	Analyse multivariée
		Densité d'individus par pièce					
Indice de niveau économique	*Faible*	**< 1.5**	164	4%	18.9	1	1
		Entre 1.5 et 2.5	407	9%	13.5	0.67 (0 41-1 08)	0.64 (0.26-1 60)
		> 2.5	794	18%	18.8	0.99 (0 64-1 53)	0 93 (0 34-2 55)
	Moyen	**< 1.5**	283	6%	10 2	1	1
		Entre 1.5 et 2.5	648	14%	11 6	1 15 (0 73-1 80)	1 18 (0 48-2 89)
		> 2.5	580	13%	13 6	1.38 (0 88-2 18)	1 52 (0 55-4 17)
	Elevé	**< 1.5**	513	11%	7.2	1	1
		Entre 1.5 et 2.5	783	17%	12 5	1 84 (1 24-2 73)	1 63 (1 08-2 46)
		> 2.5	305	7%	9 2	1 30 (0 79-2 17)	1 16 (0 68-1 97)
		Sanitaires					
Mère scolarisée	*Non*	**Privés**	1794	40%	14 4	1	1
		Autres	1343	30%	14.3	0 99 (0.80-1.22)	0.67 (0 33-1.37)
	Oui	**Privés**	926	21%	8.2	1	1
		Autres	414	9%	13 0	1 67 (1 16-2 43)	1.20 (0 79-1 83)
		Quantité d'eau disponible					
Allocations familiales	*Oui*	**A volonté**	253	6%	5 5	1	1
		Suffisante	227	5%	11 4	2.21 (1.12-4 34)	2 27 (1 13-4 58)
		Insuffisante	106	2%	17 9	3 73 (1 79-7 56)	2 45 (1.08-5.57)
	Non	**A volonté**	1300	29%	12 4	1	1
		Suffisante	1546	35%	13 2	1 07 (0 86-1 34)	0 96 (0 27-3 47)
		Insuffisante	1045	23%	15 0	1 25 (0 97-1 60)	0 83 (0 20-3 34)
		Age					
Taille de la mère	*Moins de 1 60 m*	**De 0 à 11 mois**	271	6%	7 0	1	1
		De 12 à 23 mois	273	6%	24 2	4 23 (2 45-7 28)	9 02 (2 35-34 67)
		De 24 à 35 mois	273	6%	19 0	3 12 (1 78-5 46)	6 95 (1 82-26 45)
		De 36 à 47 mois	253	6%	20 9	3 51 (2 02-6 11)	6 82 (1 80-25 87)
		De 48 à 59 mois	243	5%	22 2	3 78 (2 16-6 63)	7 69 (1 97-29 95)
	Entre 1 60 et 1 65 m	**De 0 à 11 mois**	273	6%	6 6	1	1
		De 12 à 23 mois	296	7%	16 9	2 88 (1 63-5 08)	5 78 (1 47-22 72)
		De 24 à 35 mois	288	6%	12 1	1 96 (1 08-3 54)	4 22 (1 08-16 54)
		De 36 à 47 mois	264	6%	21 6	3 90 (2 23-6 82)	7 50 (1 95-28 84)
		De 48 à 59 mois	272	6%	8 8	1 37 (0 72-2 60)	2 76 (0 68-11 26)
	Plus de 1 65 m	**De 0 à 11 mois**	339	8%	5 0	1	1
		De 12 à 23 mois	369	8%	8 1	1 68 (0 91-3 10)	3 14 (1 40-7 04)
		De 24 à 35 mois	372	8%	9 7	2 03 (1 12-3 68)	4 06 (1 84-8 97)
		De 36 à 47 mois	362	8%	10 8	2 29 (1 27-4 13)	4 06 (1 84-8 98)
		De 48 à 59 mois	329	7%	9 4	1 97 (1 07-3 63)	3 64 (1 61-8 23)
		Age					
Sexe de l'enfant	*Fille*	**De 0 à 11 mois**	399	9%	3 3	1	1
		De 12 à 23 mois	474	11%	16 0	5 67 (2 16-7 24)	3 14 (1 40-7 04)
		De 24 à 35 mois	468	10%	13 2	4 53 (2 45-8 38)	4 06 (1 84-8 97)
		De 36 à 47 mois	448	10%	15 6	5 50 (2 99-10 11)	4 06 (1 84-8 98)
		De 48 à 59 mois	428	10%	12 1	4 11 (2 20-7 66)	3 64 (1 61-8 23)
	Garçon	**De 0 à 11 mois**	484	11%	8 5	1	1
		De 12 à 23 mois	464	10%	15 1	1 92 (1 27-2 89)	1 06 (0 27-4 07)
		De 24 à 35 mois	465	10%	13 1	1 63 (1 07-2 48)	1 37 (0 36-5 25)
		De 36 à 47 mois	431	10%	18 3	2 42 (1 61-3 64)	1 76 (0 46-6 68)
		De 48 à 59 mois	416	9%	13 7	1 71 (1 12-2 63)	1 53 (0 39-6 02)
		Dernière diarrhée récente					
Sanitaires	*Privés*	**Oui**	705	16%	13 0	1 09 (0 85-1 41)	1 06 (0 80-1 39)
		Non	2015	45%	12 1	1	1
	Autres	**Oui**	516	12%	20 0	1 91 (1 45-2 53)	2 01 (1 20-3 35)
		Non	1241	28%	11 5	1	1

8b. Comparaison des Risques relatifs ajustés

Facteur modificateur d'effet		Facteur de risque	n		Prévalence %	Risque relatif ajusté ($IC_{95\%}$) Analyse trivariée		Analyse multivariée	
		Densité d'individus par pièce							
Indice de niveau économique	*Faible*	**< 1.5**	164	4%	18.9	1		1	
		Entre 1.5 et 2.5	407	9%	13.5	0 71	(0.48-1 07)	0 69	(0 32-1.50)
		> 2.5	794	18%	18.8	0 99	(0.70-1.41)	0 93	(0 39-2 21)
	Moyen	**< 1.5**	283	6%	10 2	1		1	
		Entre 1.5 et 2.5	648	14%	11.6	1 13	(0.75-1 69)	1 11	(0 51-2 43)
		> 2.5	580	13%	13.6	1 32	(0.89-1 99)	1 38	(0 57-3 35)
	Elevé	**< 1.5**	513	11%	7.2	1		1	
		Entre 1.5 et 2.5	783	17%	12 5	1 73	(1.62-1.85)	1.53	(1 07-2.20)
		> 2.5	305	7%	9.2	1.27	(0 80-2.04)	1 16	(0 73-1 86)
		Sanitaires							
Mère scolarisée	*Non*	**Privés**	1794	40%	14 4	1		1	
		Autres	1343	30%	14.3	0 99	(0.83-1 18)	0 72	(0 40-1 31)
	Oui	**Privés**	926	21%	8 2	1		1	
		Autres	414	9%	13 0	1 59	(1.14-2 21)	1 21	(0 85-1 72)
		Quantité d'eau disponible							
Allocations familiales	*Oui*	**A volonté**	253	6%	5 5	1		1	
		Suffisante	227	5%	11 4	2 07	(1 11-3 86)	1 99	(1 08-3 67)
		Insuffisante	106	2%	17 9	3 24	(1.69-6.22)	2 28	(1 16-4 52)
	Non	**A volonté**	1300	29%	12.4	1		1	
		Suffisante	1546	35%	13.2	1 06	(0.87-1.30)	0.96	(0 31-2 99)
		Insuffisante	1045	23%	15 0	1 21	(0.99-1 48)	0 88	(0.27-2 91)
		Age							
Taille de la mère	*Moins de 1.60 m*	**De 0 à 11 mois**	271	6%	7.0	1		1	
		De 12 à 23 mois	273	6%	24.2	3 45	(2 12-5 60)	6 91	(2 03-23.50)
		De 24 à 35 mois	273	6%	19.0	2.72	(1.65-4 47)	5 63	(1 67-18 98)
		De 36 à 47 mois	253	6%	20 9	2.99	(1.81-4 93)	5 31	(1.59-17 78)
		De 48 à 59 mois	243	5%	22 2	3.17	(1.93-5 19)	5 98	(1 74-20 51)
	Entre 1 60 et 1 65 m	**De 0 à 11 mois**	273	6%	6.6	1		1	
		De 12 à 23 mois	296	7%	16 9	2 56	(1 53-4 29)	4 89	(1 41-17 03)
		De 24 à 35 mois	288	6%	12 1	1 84	(1 07-3 17)	3 72	(1 07-12 93)
		De 36 à 47 mois	264	6%	21 6	3 27	(1 97-5 44)	5 88	(1 73-19 96)
		De 48 à 59 mois	272	6%	8 8	1 34	(0 74-2 41)	2 59	(0 72-9 37)
	Plus de 1 65 m	**De 0 à 11 mois**	339	8%	5 0	1		1	
		De 12 à 23 mois	369	8%	8 1	1 62	(0 91-2 88)	2 98	(1 41-6 27
		De 24 à 35 mois	372	8%	9 7	1 93	(1 10-3 37)	3 61	(1 74-7 52)
		De 36 à 47 mois	362	8%	10 8	2 15	(1.24-3 72)	3 68	(1 77-7 66)
		De 48 à 59 mois	329	7%	9 4	1 88	(1 06-3 33)	3 35	(1 58-7 13)
		Age							
Sexe de l'enfant	*Fille*	**De 0 à 11 mois**	399	9%	3 3	1		1	
		De 12 à 23 mois	474	11%	16 0	4 92	(2 77-8 72)	2 98	(1 41-6 27)
		De 24 à 35 mois	468	10%	13 2	4 06	(2 27-7 28)	3 62	(1 74-7 52)
		De 36 à 47 mois	448	10%	15 6	4 79	(2.69-8 53)	3 68	(1 77-7 66)
		De 48 à 59 mois	428	10%	12.1	3 73	(2.06-6 74)	3 35	(1 58-7 13)
	Garçon	**De 0 à 11 mois**	484	11%	8 5	1		1	
		De 12 à 23 mois	464	10%	15 1	1 78	(1 23-2 57)	1 05	(0 30-3 66)
		De 24 à 35 mois	465	10%	13 1	1 55	(1 06-2 26)	1 31	(0 37-4 58)
		De 36 à 47 mois	431	10%	18 3	2 16	(1 51-3 10)	1 67	(0 48-5 79)
		De 48 à 59 mois	416	9%	13 7	1 62	(1 10-2 37)	1 47	(0 41-5 23)
		Dernière diarrhée récente							
Sanitaires	*Privés*	**Oui**	705	16%	13 0	1 08	(0 86-1 35)	1 06	(0 85-1 32)
		Non	2015	45%	12 1	1		1	
	Autres	**Oui**	516	12%	20 0	1 73	(1 37-2 18)	1 74	(1 16-2 62)
		Non	1241	28%	11 5	1		1	

Annexe 9 : Evaluation de la prise en compte des facteurs de confusion

9a. Comparaison visuelle à partir des Odds-ratios

Facteur de risque	Odds-ratios ($IC_{95\%}$)		
	Brut	**Ajusté à un facteur de confusion**	**Ajusté à tous les facteurs du modèle**
Source d'eau		*Indice de niveau économique*	
Privée	1	1	1
Autre	1.46 (1 22-1.76)	1.14 (0.92-1 41)	1.32 (0 94-1 86)
Secteur d'activité du chef de ménage		*Niveau scolaire du chef de menage*	
Etat	0.67 (0 50-0 89)	0.80 (0.60-1 07)	0 75 (0 55-1 29)
Privé	1	1	1
Informel	1 06 (0 86-1 31)	0 95 (0 77-1 19)	0 94 (0 74-1 20)
Agriculture/autre	1.24 (0 96-1 58)	1 13 (0 88-1 46)	1 05 (0 79-1 39)
Age du chef de ménage		*Niveau scolaire du chef de menage*	
		1	
		1.10 (0 92-1 02)	
		Durée de résidence du chef de ménage à Pikine	
Moins de 50 ans	1	1	1
50 ans et plus	1.23 (1 03-1 46)	1.19 (0.99-1 42)	1 14 (0 91-1 42)
		Parente de la mère avec le chef de menage	
		1	
		1 18 (0 98-1 42)	

Remarque :
On n'a pas calculé d'indices d'association pour les facteurs de risque ayant ou étant eux-mêmes facteurs modificateur d'effet

9b. Comparaison visuelle à partir des Risques relatifs

Facteur de risque	Risques relatifs($IC_{95\%}$)		
	Brut	**Ajusté à un facteur de confusion**	**Ajusté à tous les facteurs du modèle**
Source d'eau		*Indice de niveau économique*	
Privée	1	1	1
Autre	1 39 (1.19-1 62)	1 11 (0 92-1.34)	1 22 (0 93-1 59)
Secteur d'activité du chef de ménage		*Niveau scolaire du chef de ménage*	
Etat	0.70 (0.54-0.90)	0 82 (0 63-1 07)	0 80 (0 61-1.04)
Privé	1	1	1
Informel	1.05 (0.87-1 27)	0 96 (0 80-1 16)	0 94 (0.77-1 14)
Agriculture/autre	1 20 (0.97-1 48)	1 11 (0 90-1 38)	1 06 (0 86-1 33)
Age du chef de ménage		*Niveau scolaire du chef de menage*	
		1	
		1 09 (0.93-1 27)	
		Durée de résidence du chef de ménage à Pikine	
Moins de 50 ans	1	1	1
50 ans et plus	1 20 (1 03-1 39)	1.16 (0 99-1 36)	1 13 (0 94-1.35)
		Parenté de la mère avec le chef de ménage	
		1	
		1 15 (0 98-1 36)	

Annexe 10 : Les procédures SAS utilisées pour les modèles logistique et log-binomial

les procédures SAS ci-dessous correspondent au modèle logistique. Pour le modèle log-binomial il suffit de remplacer link=logit par link=log.

L'estimation de ϕ se fait en ajoutant l'option **'dscale'**. à la suite de link=logit ou link=log.

Légende : - en gras la syntaxe SAS
- en italique les paramètres à entrer

Exemple de Tableau en entrée (données individuelles)

Observations	**Retard de taille**	**Facteurs de risque**						
		n	X_1	X_2		X_j		X_q
1	y_1	1	x_{11}	x_{21}	...	x_{j1}	...	x_{q1}
2	y_2	1	x_{12}	x_{22}	...	x_{j2}	...	x_{q2}
...	...	...	...	...	...	...	...	...
i	y_i	1	x_{1i}	x_{2i}	...	x_{ji}	...	x_{qi}
...	...	...	...	...	...	...	...	...
n	y_n	1	x_{1n}	x_{2n}	...	x_{jn}	...	x_{qn}

- modèles logistique et log-binomial avec un facteur de risque

Proc genmod data=*tableau en entrée*;
Class *facteur de risque*;
Model *retard de taille* / n = *facteur de risque* / **dist**=*binomial* **link**=*logit* **type***1*;
Run;

- modèles logistique et log-binomial avec un facteur de risque et un facteur modificateur d'effet

Proc genmod data=*tableau_en _entrée*;
Class *facteur de risque facteur modificateur d'effet*;
Model *retard de taille* / n = *facteur de risque facteur modificateur d'effet*
*facteur de risque * facteur modificateur d'effet* / **dist**=*binomial* **link**=*logit* **type***3*;
Run;

- modèles logistique et log-binomial avec un facteur de risque et un facteur de confusion

Proc genmod data=*tableau_en _entrée*;
Class *facteur de risque facteur de confusion*;
Model *retard de taille* / n = *facteur de risque facteur de confusion*/ **dist**=*binomial* **link**=*logit* **type***1* **type***3*;
Run;

- modèles logistique et log-binomial avec plusieurs facteurs

Proc genmod data=*tableau_en _entrée*;
Class *facteurs de risque facteurs de confusion facteurs modificateur d'effet*;
Model *retard de taille* / n = *facteurs de risque facteur de confusion facteurs modificateur d'effet*
*facteur de risque * facteur modificateur d'effet*
/ **dist**=*binomial* **link**=*logit* **type***1* **type***3*;
Run;

Annexe 11 : Comparaison sorties GENMOD / sorties MACROS :

Une sortie GENMOD classique :

The GENMOD Procedure

Model Information

Description	Value	Label
Data Set	WORK.TRAVAIL	
Distribution	BINOMIAL	
Link Function	LOGIT	
Dependent Variable	HAZ2	retard de taille / -2
Dependent Variable	N	
Observations Used	4477	
Number Of Events	581	
Number Of Trials	4477	

Class Level Information

Class	Levels	Values
VACCI	2	non |oui
AGECM_D2	2	>=50 |<50

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4473	3446 4153	0.7705
Scaled Deviance	4473	3446 4153	0.7705
Pearson Chi Square	4473	4476.9998	1 0009
Scaled Pearson X2	4473	4476.9998	1.0009
Log Likelihood		1723 2076	

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	-2.0033	0 0685	855 7304	0.0001
VACCI	non		1	0.0402	0.1482	0 0735	0.7863
VACCI	|oui		0	0.0000	0.0000		
AGECM_D2	>=50		1	0.1465	0.1015	2.0836	0 1489
AGECM_D2	|<50		0	0 0000	0 0000		
VACCI*AGECM_D2	non	>=50	1	0 2751	0 2138	1 6554	0.1982
VACCI*AGECM_D2	non	|<50	0	0 0000	0.0000		
VACCI*AGECM_D2	|oui	>=50	0	0.0000	0 0000		
VACCI*AGECM_D2	|oui	|<50	0	0 0000	0 0000		
SCALE			0	1 0000	0.0000		

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	3455 8646	0		
VACCI	3453.4962	1	2.3684	0.1238
AGECM_D2	3448.0705	1	5.4257	0.0198
VACCI*AGECM_D2	3446 4153	1	1.6552	0 1982

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
VACCI	1	2.7008	0.1003
AGECM_D2	1	7 0374	0 0080
VACCI*AGECM_D2	1	1.6552	0 1982

Sortie de la macro %or_rr_brut :

Exemple d'appel : %or_rr_br(data=toutpik, maladie=haz2, listfac=vacci agecm_d2 diarr wc_c2,alpha=0.05)

EXPO	MODA	N	%	PREVAL	OR_b	[$IC_{95\%}$]		RR_b	[$IC_{95\%}$]	
VACCI	non	916	20%	14.52	1.18	0.96	1.45	1.15	0.96	1.38
VACCI	|oui	3561	80%	12.58	1.00	1.00	1.00	1.00	1.00	1.00
AGECM_D2	>=50	1905	43%	14.33	1.23	1.03	1.46	1.20	1.03	1.39
AGECM_D2	|<50	2572	57%	11.98	1.00	1.00	1.00	1.00	1.00	1.00
DIARR	oui	1221	27%	15.97	1.41	1.17	1 70	1.35	1.15	1.58
DIARR	|non	3256	73%	11.86	1.00	1.00	1.00	1.00	1 00	1 00
WC_C2	wc_a	1757	39%	14.00	1.16	0.97	1.38	1.14	0.98	1.33
WC_C2	|wc_	2720	61%	12.32	1.00	1.00	1 00	1.00	1.00	1.00

Sortie de la macro %interac2 :

Exemple d'appel : %interac2(data=toutpik, maladie=haz2, listfac=vacci agecm_d2 diarr wc_c2, ass=OR, maxite=50, converg=1E-4)

OBS	FACTEUR1	FACTEUR2	CHISQ	PVALC
1	diarr	wc_c2	8.4552	0.0036
2	vacci	wc_c2	2.2164	0.1366
3	vacci	agecm	1.6552	0 1982
4	vacci	diarr	0.1436	0.7047
5	agecm	wc_c2	0 1325	0.7158
6	agecm	diarr	0.0936	0.7596

Sortie de la macro %comparor :

Exemple d'appel : %or_rr_br(data=toutpik, maladie=haz2, listfac=vacci agecm_d2 diarr wc_c2, listconf=vacci agecm_d2 diarr wc_c2,alpha=0.05, ass=OR)

EXPO	MODA_EXP	N	CONFUS	ORBRUT	[$IC_{95\%}$]		ORAJUS	[$IC_{95\%}$]	
VACCI									
VACCI	non	916	agecm_d2	1.18	0.9579	1.4544	1.18	0.9606	1.4590
VACCI	|oui	3561	agecm_d2	1.00	1.0000	1.0000	1.00	1.0000	1.0000
AGECM_D2									
AGECM_D2	>=50	1905	diarr	1.23	1.0323	1.4646	1.22	1.0257	1.4561
AGECM_D2	|<50	2572	diarr	1.00	1.0000	1.0000	1.00	1.0000	1.0000
AGECM_D2									
AGECM_D2	>=50	1905	wc_c2	1.23	1.0323	1.4646	1.27	1.0637	1.5182
AGECM_D2	|<50	2572	wc_c2	1.00	1.0000	1.0000	1.00	1.0000	1.0000
DIARR									
DIARR	oui	1221	wc_c2	1.41	1.1728	1.7026	1.41	1.1662	1.6936
DIARR	|non	3256	wc_c2	1.00	1.0000	1.0000	1.00	1.0000	1.0000
WC_C2									
WC_C2	wc_autr	1757	diarr	1.16	0 9713	1.3832	1 15	0.9594	1 3673
WC_C2	|wc_privé	2720	diarr	1.00	1.0000	1 0000	1 00	1 0000	1 0000

Annexe 12 : Calcul de mesures de risque en présence de facteurs modificateurs d'effet

Etat de santé : Retard de taille (Variable réponse)
Facteur de risque : Densité d'individus par pièce
Facteur modificateur d'effet : Indice de niveau économique
} Variables explicatives

La méthode est aussi valable pour le calcul des RR. On suppose ici un modèle à deux facteurs.

Les données nécessaires :

Les paramètres estimés par le modèle logistique

Facteurs	Modalité facteur de risque	Modalité facteur modificateur d'effet	Paramètres estimés	Erreur standard (std)
RAPP_D3	1 5-2.5	-	β1=0 610	0.202
RAPP_D3	> 2.5	-	β2=0 262	0.262
RAPP_D3	< 1.5	-	β3=0	0
RAPP_D3*ECO1_C3	1.5-2.5	Niveau1	β4=-0.473	0.307
RAPP_D3*ECO1_C3	1 5-2.5	Niveau2	β5=-1.010	0 319
RAPP_D3*ECO1_C3	1 5-2 5	Niveau3	β6=0	0
RAPP_D3*ECO1_C3	> 2.5	Niveau1	β7=0 060	0 349
RAPP_D3*ECO1_C3	> 2 5	Niveau2	β8=-0.272	0 341
RAPP_D3*ECO1_C3	> 2.5	Niveau3	β9=0	0
RAPP_D3*ECO1_C3	< 1 5	Niveau1	β10=0	0
RAPP_D3*ECO1_C3	< 1.5	Niveau2	β11=0	0
RAPP_D3*ECO1_C3	< 1 5	Niveau3	β12=0	0

Matrice des variances covariances des paramètres du modèle

	β1	β2	β3	β4	β5	β6	β7	β8	β9	β10	β11	β12
β1	0 041											
β2	0.029	0 068										
β3	0	0	0									
β4	-0 041	-0 029	0	0.094								
β5	-0.041	-0.029	0	0.041	0 101							
β6	0	0	0	0	0	0						
β7	-0 029	-0 068	0	0 067	0 029	0	0 121					
β8	-0 029	-0 068	0	0 029	0 068	0	0 068	0 116				
β9	0	0	0	0	0	0	0	0	0			
β10	0	0	0	0	0	0	0	0	0	0		
β11	0	0	0	0	0	0	0	0	0	0	0	
β12	0	0	0	0	0	0	0	0	0	0	0	0

Les formules utilisées :

$$OR_{\left(\frac{Niveau1}{1\,5-2\,5}_{<1\,5}\right)} = e^{(\beta 1+\beta 4)} \qquad IC_{95\%} = e^{\left(\beta 1+\beta 4\pm 1\,96\sqrt{std1^2+std4^2+2\,\mathrm{cov}(\beta 1,\beta 4)}\right)}$$

$$OR_{\left(\frac{Niveau1}{1\,5-2\,5}_{<1\,5}\right)} = e^{(0\,610-0\,473)} = 1.147 \qquad IC_{95\%} = e^{\left(0\,610-0\,473\pm 1\,96\sqrt{0\,202^2+0\,307^2+2*-0\,041}\right)} = [0.730\,;1.801]$$

Les résultats obtenus :

Indice de niveau économique	Densité d'individus par pièce	OR_a	IC 95%	
Niveau1	< 1.5	1	1	1
	1.5-2 5	1.147	0.730	1.801
	> 2.5	1.380	0.873	2.180
Niveau2	< 1.5	1	1	1
	1 5-2.5	0.670	0.266	1.687
	> 2 5	1.706	1.106	2.631
Niveau3	< 1.5	1	1	1
	1 5-2 5	1.840	1.239	2.734
	> 2.5	1.299	0.778	2.172

Annexe 13 : Tests de Hosmer et Lemeshow pour les modèles logistique et log-binomial

Groupe	Modèle logistique			Modèle log-binomial		
	Total	Observés	Théoriques	Total	Observés	Théoriques
1	447	13	10,37	447	14	11,23
2	448	23	20,34	448	23	21,22
3	448	27	28,66	448	28	29,24
4	448	36	36,03	448	31	36,33
5	447	38	43,61	447	46	43,24
6	448	43	52,68	448	42	51,89
7	448	71	64,44	448	69	62,91
8	448	85	79,68	448	81	77,47
9	448	92	101,25	448	92	99,18
10	447	153	143,92	447	155	149,33
	Khi²=	6.05		**Khi²=**	5.21	
	P-value=	0.64		**P-value=**	0.73	
	Ddl=	8		**Ddl=**	8	

Les observés correspondent au nombre d'enfants ayant un retard de taille observés dans chaque groupe. Les théoriques sont les effectifs prédits par le modèle et s'obtiennent pour chaque classe en faisant le produit de la probabilité moyenne de retard de taille de la classe avec son effectif.

Annexe 14 : La procédure GENMOD dans les différentes étapes de l'analyse

GENMOD est la procédure SAS utilisée pour écrire des modèles linéaires généralisés. Elle a été utilisée à tous les stades de l'analyse :

- Calcul des paramètres β permettant d'estimer les OR et RR bruts ainsi que les Intervalles de confiance à 95%

- Recherche des facteurs modificateurs d'effet et des facteurs de confusion par tests de rapports de vraisemblance

- Mise en place du modèle logistique et du modèle log-binomial

- Calcul des paramètres β et de leur matrice de covariances[14] permettant d'estimer les OR et RR ajustés ainsi que les intervalles de confiance à 95%

- Estimation du paramètre de dispersion ϕ (ajustement des modèles en cas de surdispersion)

- Estimation des probabilités d'apparition du retard de taille pour chaque individu (test de Hosmer et Lemeshow)

[14] Nécessaire pour l'ajustement des indices d'association sur les facteurs modificateurs d'effet (**Annexe 12**)

Annexe 15 : La macro %or_rr_brut

```
/*********************************************************************
MACRO or_rr_br - Utilisation de proc freq et proc genmod

Thierry HOARAU Dakar juin 99

Cette macro permet de calculer les OR et RR brut pour
une liste de facteurs de risque

Conditions d'application

-classes de référence préalablement définies si nécessaire
-variable réponse codée en 1/2 (1 malade ; 2 non malade)

Paramètres

 %MACRO or_rr_br(DATA=, MALADIE=, LISTFAC=,ALPHA=0.05, OUT=TOTO),

-DATA=tableau en entrée
-MALADIE=variable codant l'état nutritionnel (1/2)
-LISTFAC=liste de facteurs de risque potentiels
-ALPHA=risque d'erreur des intervalles de confiance
        * 0 05 (par défaut)
        * 0 01 (ALPHA=0 01)
 OUT=tableau en sortie (toto par défault)

Remarques sur les sorties

 OR et RR brut
-Intervalles de confiance associés (0 05 par défaut)
 Prévalence et effectif par classe
 tests de rapport de vraisemblance (khi2 et ddl) testant la
significativité du modèle

Exemple d_appel voir fin

*********************************************************************/

title'',

%MACRO or_rr_br( DATA=, MALADIE= LISTFAC=,ALPHA=0 05, OUT=TOTO),

 %let BORN=1 96,
 %if &ALPHA 0 01 %then %let BORN=2 576

data &out,
run,

/*Construction de la table de travail*/

DATA TRAVAIL,
 set &DATA (keep=&MALADIE &LISTFAC),
 n=1,
 &MALADIE=2 &MALADIE,
 format &MALADIE,
RUN;


/*Début de la boucle*/

%let nbfac=1,
%let fac=%scan &LISTFAC &nbfac.,
%do %while'&fac^=
```

```
proc freq data=TRAVAIL ,
tables (&fac)* &maladie / out=nn outpct  ;
run;

data _null_;
 set NN;
 file "c \test.asc",
 put &fac &maladie count percent pct_row;
run,
data titi;
 infile "c:\test.asc" ,
 length moda_exp $14 ,
 input moda_exp &maladie n percent preval;
run,
data titi;
 set titi;
 where &maladie=1,
 total=n*100/percent;
 nrow=n*100/preval ,
 perc=nrow/total;
run;

PROC GENMOD DATA=TRAVAIL;
   class &fac ;
   title "Calcul de l'OR brut de &fac ";
   model &MALADIE / n = &fac /dist=binomial link=logit  TYPE1;
   make 'parmest' out=parmor ,
   make 'TYPEI' out=typelor ;

RUN;

DATA _NULL_;
file log;
put "Risque relatif brut de la variable &fac",

  PROC GENMOD DATA=TRAVAIL,
   class &fac ;
   title "Calcul de RR brut de &fac ";
   model &MALADIE / n = &fac /dist=binomial link=log TYPE1,
   make 'parmest' out=parmrr ,
   make 'type1' out=type1rr ,

RUN;
/*Table contenant Khi2 associé à l'OR brut*/

DATA TYPE1OR;
 set typelor(drop = DEV),
 where SOURCE="%upcase(&fac)",
 rename DF=DF_OR;
 rename SOURCE=EXPO;
 rename CHISQ=X2or,
 rename PVALC=Por;
RUN;

/*Table contenant Khi2 associé au RR brut*/

DATA TYPE1rr,
 set type1rr (drop = DEV),
 where SOURCE="%upcase(&fac)",
 rename DF=DF_rr;
 rename SOURCE=EXPO,
 rename CHISQ=X2rr;
 rename PVALC=Prr,
RUN,

DATA PARMOR,
 set parmor (drop = chisq pval df),
```

```
 where parm="%upcase(&fac)"
 rename parm=expo,
 rename level1=moda_exp,
 rename estimate=estor,
 rename stderr=stdor,
RUN,

DATA PARMRR,
 set parmrr (drop = chisq pval df);
 where parm="%upcase(&fac)",
 rename parm=expo,
 rename level1=moda_exp,
 rename estimate=estrr,
 rename stderr=stdrr,
RUN,

/* Tri de parmor et de parmrr en vue de leur fusion */

 PROC SORT data=tit1
  by moda_exp
 RUN,
 PROC SORT DATA=parmor,
  by moda_exp,
 RUN,
 PROC SORT DATA=parmrr,
  by moda_exp,
 RUN,

/*concaténation verticale avec changement des noms de variables */

DATA ORRR,
   merge parmor
         parmrr
   ,
   by moda_exp ,
   where (expo="%upcase(&fac)" ., /*pour avoir uniquement les lignes*/
   OR=exp(estor),           /*intéressantes*/
   RR=exp(estrr),
   ICor_INF=exp(estor  &BORN *stdor),
   ICor_SUP=exp(estor +&BORN *stdor),
   ICrr_INF=exp(estrr  &BORN *stdrr),
   ICrr_SUP=exp(estrr +&BORN *stdrr);
RUN,

DATA ORRR2,
  merge orrr (drop = stdrr stdor estor estrr)
        tit1 (drop = &maladie n percent total ),
  by moda_exp,
RUN,

DATA ORRR3,
  merge orrr2
        typelor
        typelrr
RUN,

DATA &out
 set &out orrr3,
RUN,
DATA &out,
set &out,
where expo<>"",
run,

proc datasets lib=work,
 delete typelor typelrr orrr orrr2 orrr3 nn parmor parmrr tit1 ,
 run,
```

```
%let nbfac=%eval(&nbfac+1),
%let fac=%nrbquote(%scan(&listfac,&nbfac)),
%end;

/* fin de la boucle*/

/*dm'clear output';  */

 PROC PRINT DATA=&out noobs;
   options nodate,
   TITLE1 "Comparaison entre OR et RR brut",
   title2 "Maladie = &MALADIE",
   title3 "------",
   TITLE4 "Fichier de travail . &data",
   title6 "------",
   TITLE8 "Risque d_erreur des Intervalles de confiance = &ALPHA";
   VAR expo moda_exp nrow perc  preval OR ICor_INF  ICor_SUP  RR ICrr_INF ICrr_SUP
   x2or por x2rr prr,
   format OR  RR  preval  x2or x2rr  ICor_INF  ICor_SUP  ICrr_INF ICrr_SUP 5 2,
   format  perc  percent ;
   format nrow 5.;
 RUN;
options noxwait,
x 'del c:\test.asc',

 %MEND or_rr_br;

/*Exemple d appel

%let var= nbpie_c3 wc_c2 eau_c2 oqot1 ,
%or_rr_br(DATA=toutpik nonmanq2, MALADIE=haz2, LISTFAC=&var);

*/
```

Annexe 16 : La macro %interac2

```
/*************************************************************************
MACRO INTERAC2   Utilisation de proc genmod

AVANT PROPOS
---------- --
INTERAC2 n'est pas une version plus récente de INTERAC1  La différence
entre ces 2 macros se situe sur les paramètres en entrée
 -interac1 applicable pour une liste de facteurs de risque distincte
 de celle des facteurs d'interaction
 -interac2 applicable lorsque la liste est la même pour risque et
 interaction
--- -                        -----   ---    ----   -  - --
Thierry HOARAU Dakar avril 99
*

Cette macro permet de déterminer si un facteur est d'interaction ou non
 par test de rapport de vraisemblance (apport de l'interaction dans le
 le modèle avec un facteur de risque et un facteur d'interaction potentiel)

Conditions d'application

-données manquantes ou pas (table automatiquement nettoyée ici)
-classes de référence préalablement définies si nécessaire
-variable réponse codée en 1/2

Paramètres

 %MACRO INTERAC2(DATA=, MALADIE=, LISTFAC= ,ASS=OR ,
                 MAXITE=50 ,CONVERG=1e-4, OUT=TOTO),

-DATA=tableau en entrée
-MALADIE=variable codant l'état nutritionnel (1/2)
 LISTFAC=liste de facteurs de risque et de confusion potentiels
 ASS=indice d'association utilisé
        *OR   odds-ratio (par défaut)
        *RR   risque relatif (ASS=RR)
-MAXIT=le maximum d'itérations que genmod doit effectuer pour trouver
       le max de vraissemblance (l'augmenter si pb de convergence)
       (par défaut = 50)
-CONVERG=la procédure itérative s'arrête lorsque l'écart entre deux max
         de vraissemblance est égal à cette valeur
         (par défaut = 1e-4) (le dimunuer si pb de convergence)
-OUT=tableau en sortie (toto par défault)
Remarque    Attention à l'utilisation de maxit et de converge  Les faire trop
            varier peut jouer sur la significativité du modèle
            Une solution de dernier recours existe si le problème de convergence
            persiste   option INITIAL (compliqué)

L'interprétation se fait sur un test du khi2 comparant le modèle sans et
le modèle avec interaction   (type3)

Exemple d'appel voir fin

*************************************************************************/
/*
%let DATA=toutpik nonmanq2,
%let MALADIE=haz2;
%let LISTFAC=vacci wc_c2,
%let ASS=OR,
%let  MAXITE=50;
%let CONVERG=1E-4;
%let OUT=TOTO;
%let fac=vacci;
%let int=wc_c2,
*/

%MACRO INTERAC2(DATA=, MALADIE=, LISTFAC=,ASS=OR ,
                MAXITE=50, CONVERG=1E-4, OUT=TOTO),

 %let LINK=LOGIT,
 %if %nrbquote(%upcase(&ASS))=RR %then %let LINK=LOG;

data &out,
run;

/*nettoyage de la table*/

DATA TRAVAIL,
 set &DATA (keep=&MALADIE &LISTFAC ),
 n=1,
 format &MALADIE,
 &MALADIE=2-&MALADIE,
 array tab &MALADIE &LISTFAC ,
 do over tab,
 if tab=. then delete;
 end,
RUN;


/*Début de la double boucle*/

%let nbfac=1,
%let fac=%scan(&LISTFAC,&nbfac);
%do %while(&fac^=),
   %let nbintr=1,
   %let int=%scan(&LISTFAC,&nbintr);
   %do %while(&int^=);

         %if %nrbquote(%upcase(&fac)) ^= %nrbquote(%upcase(&int))
         and &nbintr>&nbfac %then %do,
         /*ne fait pas les calculs quand fac=int et LISTFAC=LISTINTR*/


     PROC GENMOD DATA=TRAVAIL /*ORDER=INTERNAL*/;
      title "&ASS . interaction &fac*&int",
      class &fac &int,
      make 'type3lr' out=type3 noprint,
      model &MALADIE/n=&fac &int &fac * &int/dist=binomial link=&LINK type3
                                              MAXIT=&MAXITE
```

```
                                        converge=&CONVERG ,
     RUN,

      DATA type3,
      set type3,
      where SOURCE="%upcase(&fac*&int)" ,
     RUN,
     DATA type3,
      set type3,
      facteur1="&fac",
      facteur2="&int",
     RUN

     DATA _null_,
      file log,
      put "Les variables croisées sont &fac et &int";
     RUN;

     DATA &out,
      set type3 &out,
     RUN,


    %end, /*fin de la boucle if conf=fac et listfac=listintr*/



    %let nbintr=%eval(&nbintr+1),
    %let int=%nrbquote(%scan(&LISTFAC,&nbintr));
  %end;
 %let nbfac=%eval(&nbfac+1);
 %let fac=%nrbquote(%scan(&listfac,&nbfac));
%end;

/* fin de la double boucle*/


PROC SORT DATA=&out,
 by pvalc,
RUN,
PROC PRINT DATA=&out,
 title1 "&ASS   Tests d_interaction entre 2 facteurs",
 title2 "résultats triés",
 title3 "Tableau en entrée : &data";
 var facteur1 facteur2 chisq pvalc;
RUN,

PROC DATASETS library=work;
 delete type3,
RUN;

 %MEND INTERAC2,

/*Exemple d'appel

%INTERAC2(DATA=toutpik.multivar, MALADIE=haz2, LISTFAC=agem_c3 elec ecol_c3,
          ASS=OR , MAXITE=50, CONVERG=1E-4, OUT=TOTO);
```

```
*/
```

Annexe 17 : La macro %comparor

```
/*************************************************************************
MACRO COMPAROR   Utilisation de proc genmod

Pierre TRAISSAC & Thierry HOARAU Dakar avril 99
*

Cette macro permet de déterminer si un facteur est de confusion ou non
 par :
  -comparaison visuelle entre OR brut et ajusté à un seul
   facteur de confusion.
  -tests du rapport de vraissemblance (tests de l'effet brut et de l'effet ajusté).

Conditions d'application

-données manquantes ou pas (table automatiquement nettoyée ici)
-classes de référence préalablement définies si nécessaire
-variable réponse codée en 1/2 (1:malade ; 2 non malade)

Paramètres

 %MACRO COMPAROR(DATA=, MALADIE=, LISTFAC=, LISTCONF= ,ASS=OR ,ALPHA=0 05, OUT=TOTO),

-DATA=tableau en entrée
-MALADIE=variable codant l'état nutritionnel (1/2)
-LISTFAC=liste de facteurs de risque potentiels
-LISTCONF=liste de facteurs de confusion potentiels (le 1er conf doit
être égal à fac '')
-ASS=indice d'association utilisé .
        *OR   odds-ratio (par défaut)
        *RR   risque relatif (ASS=RR)
-ALPHA=risque d'erreur des intervalles de confiance :
        * 0 05 (par défaut)
        * 0 01 (ALPHA=0.01)
-OUT=tableau en sortie (toto par défault)

Remarques sur les sorties :

-OR brut et ajusté pour un facteur de confusion
-Intervalles de confiance associés (0 05 par défaut)
-khi² brut et ajustés, ddl, pvalue
 sur la ligne khi2 les valeurs correspondent à :
 DF_brut  khi2_brut  pval_brut  DF_ajus  khi2_ajus  pval_ajus
-Dans le log ne pas tenir compte de :

Exemple d_appel voir fin

*************************************************************************/
title'';

%MACRO COMPAROR(DATA=, MALADIE=, LISTFAC=, LISTCONF= ,ASS=OR ,ALPHA=0 05, OUT=TOTO);

DATA &out,
RUN;

 %let BORN=1 96,
 %let LINK=LOGIT;
 %if %nrbquote(%upcase(&ASS))=RR %then %let LINK=LOG;
 %if &ALPHA=0.01 %then %let BORN=2 576,

data OOOO,
run,

/*nettoyage de la table*/

DATA TRAVAIL;
 set &DATA (keep=&MALADIE &LISTFAC &LISTCONF);
 n=1;
 &MALADIE=2-&MALADIE;
 format &MALADIE;
 array tab &MALADIE &LISTFAC &LISTCONF;
 do over tab;
 if tab=. then delete;
 end;
RUN;

/*Détermination du nombre de valeurs supprimées*/

DATA VALBRUT;
 set &DATA;
 &MALADIE=2-&MALADIE;
 n=1;
 run;
PROC GENMOD DATA=VALBRUT;
  class &LISTFAC &LISTCONF ;
  model &MALADIE / n = &LISTFAC &LISTCONF /dist=binomial link=&LINK ;
  make 'modinfo' out=modinfo noprint ;
RUN;


/*Début de la double boucle*/

%let nbfac=1;
%let fac=%scan(&LISTFAC,&nbfac);
%do %while(&fac^=);
 %let nbconf=1;
 %let conf=%scan(&LISTCONF,&nbconf);
 %do %while(&conf^=);

/*data g2;
 set g;
 length moda_exp $14. ;
 moda_exp=put(elec,elecf.);
run; */

  %if %nrbquote(%upcase(&fac)) ^= %nrbquote(%upcase(&conf)) %then %do,
  /*ne fait pas les calculs quand fac=conf*/

proc freq data=TRAVAIL ;
tables &fac / out=NN ;
run;

data _null_;
 set NN;
 file "c:\test.asc";
 put &fac 'µµµ' count;
run;
data titi;
 infile "c:\test.asc" delimiter='µµµ',
 length moda_exp $14.;
 input moda_exp n;
run;

PROC GENMOD DATA=TRAVAIL;
   class &fac ;
   title "Calcul de l'&ASS brut de &fac ";
   model &MALADIE / n = &fac /dist=binomial link=&LINK ;
   make 'parmest' out=parmbrut ;

RUN,

PROC GENMOD DATA=TRAVAIL;
```

```
   class &fac &conf ,
   title"Calcul de l'&ASS de &fac ajusté à &conf",
   model &MALADIE / n = &fac &conf /dist=binomial link=&LINK type1 type3 ,
/* sortie des ddl, khi2 et pvalue bruts et ajustés */
   make 'parmest' out=parmajus ;
   make 'TYPE1' out=type1a;
   make 'TYPE3LR' out=type3a;
RUN,

/*Table contenant Khi2 associé à l'OR brut*/

DATA TYPE1A,
 set type1a;
 where SOURCE="%upcase(&fac)";
 rename DF=&ASS_BRUT,
 rename SOURCE=EXPO;
 rename CHISQ=ICB_INF,
 rename PVALC=ICB_SUP;
 moda_exp="      Khi2       ";
 confus="-- -     ",
RUN,

/*Table contenant Khi2 associé à l'OR ajusté*/

DATA TYPE3A,
 set type3a;
 where SOURCE="%upcase(&fac)",
 rename DF=&ASS_AJUS;
 rename SOURCE=EXPO,
 rename CHISQ=ICA_INF;
 rename PVALC=ICA_SUP;
confus="&conf",
RUN;

/* Tri de parmbrut et de parmajus en vue de leur fusion */

PROC SORT data=parmbrut,
  by level1,
RUN;
PROC SORT data=parmajus,
  by level1,
RUN;
PROC SORT data=titi,
by moda_exp,
run;

/*concaténation verticale avec changement des noms de variables */

DATA ORRR;
   merge parmbrut(rename=(parm=expo level1=moda_exp estimate=estbrut
                          stderr=stdbrut) drop=chisq pval)
         parmajus(rename=(parm=expo level1=moda_exp estimate=estajus
                          stderr=stdajus) drop=chisq pval)
   ;
   by moda_exp ;
   where (expo="%upcase(&fac)" ); /*pour avoir uniquement les lignes*/
   &ASS.BRUT=exp(estbrut);           /*intéressantes*/
   &ASS_AJUS=exp(estajus),
   ICB_INF=exp(estbrut -&BORN *stdbrut),
   ICB_SUP=exp(estbrut +&BORN *stdbrut);
   ICA_INF=exp(estajus -&BORN *stdajus);
   ICA_SUP=exp(estajus +&BORN *stdajus);
   confus="&conf";
RUN,

data ORRRR;
merge titi
      ORRR ,
by moda_exp,
run;

/* Concaténation horizontale*/

DATA TYPE;
merge type1a type3a;
RUN;
DATA OOOO;
set TYPE ORRRR;
RUN;

%end; /* fin de la boucle if conf=fac*/

 DATA &out;
  set &out OOOO;
 RUN;

/*
proc datasets lib=work;
 delete type1a type3 orrr orrrr type nn oooo;
 run; */

 %let nbconf=%eval(&nbconf+1),
 %let conf=%nrbquote(%scan(&listconf,&nbconf));
%end;
%let nbfac=%eval(&nbfac+1);
%let fac=%nrbquote(%scan(&listfac,&nbfac));
%end;

/* fin de la double boucle*/

/*dm'clear output';  */

 PROC PRINT data=modinfo;
 title1 "Informations sur &DATA avant nettoyage" ;
 title2 "Déterminations des valeurs éliminées";
 RUN;

 PROC PRINT DATA=&out noobs;
   options nodate;
   TITLE1 "Table de détermination des facteurs de confusion";
   title2 "Maladie = &MALADIE";
   title3 "------";
   title5 "Table de travail : &data";
   title7 "Mesure du risque : &ass";
   title9 "------";
   TITLE10 "Risque d_erreur des Intervalles de confiance des &ASS = &ALPHA";
   VAR expo moda_exp n confus &ASS.BRUT ICB_INF  ICB_SUP  &ASS.AJUS ICA_INF ICA_SUP,
   format  ICB_INF  ICB_SUP  ICA_INF ICA_SUP 7.4;
   format &ASS.BRUT  &ASS.AJUS  5.2;
 RUN;

 /* on delete les tableaux de travail */
 proc datasets library=work;
  delete  OOOO NN ORRR valbrut orrrr PARMAJUS PARMBRUT titi TYPE TYPE1A TYPE3A;
 run;

options noxwait;

x 'del c:\test.asc';

 %MEND COMPAROR;

/*Exemple d appel

%let var= nbpie_c3 wc_c2 eau_c2 oqoti ;
%COMPAROR(DATA=toutpik.nonmanq2, MALADIE=haz2, LISTFAC=tailm_c3  , LISTCONF= merbio),

*/
```

Annexe 18 : Le programme SAS du test de Hosmer et Lemeshow

```
/* Thierry HOARAU Montpellier 27/07/99*/

/************************************
 *                                  *
 *    TEST DE HOSMER ET LEMESHOW    *
 *                                  *
 ************************************

 But du test   test d'adéquation d'un modèle (LOGISTIQUE, LOGBINOMIAL OU LOGPOISSON    ) par rapport aux
données


 2 méthodes   1) Utilisation de proc logistic option lackfit (uniquement pour la régression logistic)
              2) Application de la méthode de HM pour n'importe quel modèle (proc genmod)


 Quelques infos sur le principe du test de Hosmer et Lemeshow

        -découpe l'échantillon en 10 classes de proba d'apparition de la maladie (déciles)
        -comptage du nb de malades dans chaque classe  (effectifs observés)
        -calcul de la proba moyenne de maladie dans chaque classe multiplié par effectif classe
(effectifs théoriques)
        -test du khi2 à 8 ddl



1) PROC LOGISTIC

Le test est déjà programmé, il suffit de le demander avec l'option LACKFIT

  Méthode       logistic ne sait pas traiter des données qualitatives
                on fait donc un codage disjonctif complet grace à GLMMOD
                puis LOGISTIC et l'instruction LACKFIT

  Interprétation  · dans la rubrique "Hosmer and Lemeshow Goodness-of-Fit Test"
                    le test du khi2 compare des effectifs théoriques avec des effectifs observés
                    Ho · le modèle s'ajuste bien aux données


 Exemple de syntaxe utilisant proc logistic        */
------------ --------------------------------


%LET VARLIE=/* ******** variables menage ************ */
              rapp_d3 eau_c2 ogoti wc_c2
              ecol_c3 agecm_d2 ethc_d4 nivcm_c3 secc_c3 pikcm_d3
              parm_c3 enfmpe marim_c2 ethm_d4 scom_d2
              alloc tailm_c3 age_d5 pnais enfp_c3
              diarr singe vacci sexecm sexe;


 /*TABLE DE TRAVAIL

attention ''

-la maladie doit être codée de sorte que l'état de santé étudié ait la plus petite valeur
-de plus il faut travailler en données individuelles pour obtenir des proba par individus*/
```

```
  data travail;
   set toutpik.nonmanq2(keep=haz2 &varlie);
  run;


/*FABRICATION DE LA TABLE AVEC LES INDICATRICES */

 PROC GLMMOD DATA=travail OUTDESIGN=indics OUTPARM=parms;
                              /*indics : matrice des indicatrices*/
                              /*parms : matrice de correspondance des indicatrices
                                        avec les modalités des variables*/
   class &varlie;
   model haz2 = &varlie age_d5*sexe age_d5*tailm_c3 rapp_d3*ecol_c3
                        diarr*wc_c2 ogoti*alloc
                        wc_c2*scom_d2;
 RUN;

/* RESULTATS */

 PROC LOGISTIC data=indics;
   MODEL haz2 = col2-col116 /lackfit ;  /*pas de col1 car logistic réestime le paramètre constant*/
   output out=pred p=predi;
 run;


/*ON PEUT RETROUVER LES PROBA PREDITES GRACE A UNE INSTRUCTION OUTPUT




2) PROC GENMOD


Il va falloir programmer le test (pas difficile)

-on demande à genmod les proba estimées de haz2 (proc genmod make obstats)
-découpage en déciles (proc rank)
-calcul de la proba moy par décile (proc means by) puis des effectifs théoriques
-calcul des effectifs observés par classe
-test du khi2   */


/* préparation de la table de travail : données individuelles*/

data travail;
set toutpik.nonmanq2 (keep = haz2 &varlie);
n=1;
haz2=2-haz2;
run;


/*Calcul des proba estimées*/

PROC GENMOD DATA=TRAVAIL;
 class &varlie;
 model haz2 / n = &varlie age_d5*sexe age_d5*tailm_c3 rapp_d3*ecol_c3
                        diarr*wc_c2 ogoti*alloc
                        wc_c2*scom_d2
                /  obstats dist=binomial link=logit;
 make  'obstats' out=predic noprint;
run;

/*Classement en déciles*/

proc rank data=predic groups=10;
var pred;
ranks predi;
```

```
run;

/* Simplification de la table */

data data1;
set data1 (keep=  yvar1 /*haz2*/ predi pred );
run,

proc sort data=data1,
 by predi yvar1,
run;

/*calcul proba moy estimées par classe puis calcul des effectifs théoriques*/

proc means data=data1,
var pred ;
by predi;
output out=sort n=n mean=phaz1,
run;

data sort,
set sort;
npred=phaz1*n,
run;

/*calcul des effectifs observés*/


proc means data=data1,
var yvar1 ,
by predi yvar1,
output out=sort2 n=n ,
run,

data sort2,
set sort2,
where yvar1=1,
run;

/* fusion dans un meme tableau des effectifs observés et théoriques */

data sort3,
merge sort sort2
run;

proc freq data=sort3,
tables  predi/testf =(  10 3738749862
                        20 3412192615
                        28.6557017075
                        36.0325765764
                        43 6148971377
                        52 6836909376
                        64.4390903511
                        79.6848791318
                        101 2543660521
                        143.9197038632 );
weight  n;
output out=chi chisq;
run;


data chi,
set chi (keep=_pchi_);
pval=1-probchi(_pchi_,8),
run;

proc print;
run;
```

Annexe 19 : Les étapes de l'analyse

1- Variables explicatives retenues pour l'étude du retard de taille :

- choix basé sur les hypothèses établies d'après les études bibliographiques
- recodage des variables quantitatives et de certaines variables qualitatives (diminution du nombre de modalités) selon des critères épidémiologiques et d'équilibrage des effectifs de classe
- construction d'un indice de biens et d'un indice de niveau économique

→liste de facteurs (**38 variables**) cf **Tableau 2**

2- Analyse de la relation brute entre les facteurs de risque et le retard de taille : **analyse bivariée**

- étude qualitative (1er criblage) : tests d'indépendance du khi² (sélection des facteurs liés au retard de taille avec α=20%)
- sélection des unités statistiques ne contenant aucune valeur manquante pour les 25 variables liées au retard de taille (de n=4567 à n=4477)
- étude quantitative : calcul des OR et RR bruts et de leur intervalle de confiance (95%) pour les 25 variables sélectionnées

→liste de facteurs liés au retard de taille (**25 variables**)
→premiers résultats : OR, RR bruts et Intervalles de confiance à 95%
→comparaison des modèles logistique et log-binomial *[Ecriture d'une macro SAS]*

3- Recherche des facteurs modificateurs d'effet par analyse des interactions entre facteurs : **analyse trivariée**

Les OR (resp. RR) sont-ils égaux dans chaque strate du facteur modificateur d'effet ?

- tests de rapport de vraisemblance (seuil de signification de 5%) *[Ecriture d'une macro SAS]*

→liste de facteurs modificateurs d'effet (**11 variables**)

4- Détermination des facteurs de confusion : **analyse trivariée**

Les OR (resp. RR) bruts sont-ils égaux aux OR (resp. RR) ajustés à un facteur de confusion ?

- double test de rapport de vraisemblance (seuil de signification de 5%) *[Ecriture d'une macro SAS]*

→liste de facteurs de confusion pour information et discussion (**7 variables**)

5- Indices d'association liés à plusieurs facteurs : **analyse multivariée**

- 25 variables et 11 interactions intégrées initialement dans le modèle
- Régression pas à pas uniquement sur les interactions : 25 variables et 6 interactions retenues dans le modèle final
- Test de Hosmer et Lemeshow

→OR et RR ajustés à plusieurs facteurs
→comparaison des modèles logistique et log-binomial
→comparaison analyses bi-, tri- et multivariée

RESUME

T. HOARAU. Utilisation de modèles de régression logistique et log-binomiale dans l'étude du risque du retard de croissance en taille chez le jeune enfant sénégalais en milieu urbain.

L'identification de facteurs associés au retard de taille chez les enfants de moins de 5 ans a été menée à partir de données anthropométriques et socio-économiques collectées au cours d'une enquête nutritionnelle transversale réalisée à Pikine (Sénégal) en 1996 sur un échantillon aléatoire représentatif de 4591 enfants par l'Unité de Nutrition de l'IRD.
Les associations entre le retard de taille et les différents facteurs ont été testées et quantifiées. Les facteurs d'ajustement ont été identifiés par comparaisons de modèles emboîtés. L'utilisation de modèles linéaires généralisés tels que la régression logistique et la régression log-binomiale a permis d'estimer le risque lié à plusieurs facteurs. Les éventuelles interactions ont été prises en compte dans l'ajustement. L'adéquation des modèles aux données a été évaluée par le test de Hosmer et Lemeshow. Des macros SAS réutilisables ont été écrites.

Les modèles utilisés s'ajustent bien aux données.
L'analyse brute a montré une forte association du risque retard de taille avec l'âge de l'enfant, la taille de la mère mais aussi le niveau scolaire du chef de ménage ainsi que l'indice de niveau économique du ménage.
Après ajustement par les régressions logistique et log-binomiale, nous n'avons plus trouvé de liaison évidente du retard de taille avec le niveau scolaire du chef de ménage. L'effet de l'indice de niveau économique apparaît lié à d'autres facteurs. En revanche les associations sont maintenues avec l'âge de l'enfant et la taille de la mère. Une nouvelle association est apparue avec la quantité d'eau dont dispose le ménage.

Finalement les résultats sont discutés sur l'utilisation de l'odds-ratio comme approximation du risque relatif et sur la comparaison des résultats bruts et ajustés. Le risque relatif doit être utilisé lorsque le type d'enquête le permet. Il est également préférable d'adopter une approche multifactorielle du risque en complément d'une analyse brute.

Mots-clés : retard de croissance, odds-ratios, risques relatifs, régression logistique, régression log-binomiale, interaction, test de Hosmer et Lemeshow

ABSTRACT

Use of logistic and log-binomial regressions for studying the stunting risk of senegalese child in urban area.

This study aimed at identifying factors linked to risk of stunting. The anthropometric and socio-economic data of a representative sample of 4591 children <5 years had been obtained from a cross-sectional survey carried out in Pikine (Senegal) in 1996 by the IRD Nutrition Unit.
Relationships between stunting and risk factors were tested and quantified. Comparison of embedded models allowed identification of confounders and modifiers. Generalised linear models such as logistic regression and log-binomial regression were used for estimating adjusted measures of association. Possibly interactions were taken into account. The goodness-of-fit of the models was assessed by Hosmer and Lemeshow test. SAS macros were written.

The models provided a good description of data.
Univariate analysis showed a strong link between risk of stunting and age of children, height of mother, head of household's studying level and economic level.
After adjustment by logistic and log-binomial regressions, links between risk of stunting and head of household's studying level disappeared. The effect of the economic level appeared to be linked to others factors. However, age of children and height of mother was still linked with risk of stunting. A new association is appeared with household's available water quantity.

Finally the interpretation of odds-ratios as relative risks and the comparison of crude and adjusted results are discussed. Relative risk should be used if the survey allows it. A risk multifactor approach compared with crude analysis should be preferable.

Keywords : stunting, odds-ratios, relative risks, logistic regression, log-binomial regression, interaction, Hosmer and Lemeshow test.