COMPARAISON DE MÉTHODES POUR LA CORRECTION DE LA NON-RÉPONSE TOTALE : MÉTHODE DES SCORES ET SEGMENTATION

Émilie Dequidt[1], Benoît Buisson[2] & Nicolas Sigler[3]

[1] *Insee, Direction régionale des Pays de la Loire, Service Études Diffusion, 105 rue des Français Libres, BP 67401, 44274 NANTES CEDEX 2 ; emilie.dequidt@insee.fr*
[2] *DGFIP, Service des Retraites de l'État, 10 boulevard Gaston Doumergue, 44964 NANTES CEDEX 9 ; benoit.buisson@dgfip.finances.gouv.fr*
[3] *Insee, Direction régionale des Pays de la Loire, Pôle Ingénierie statistique Entreprises, 105 rue des Français Libres, BP 67401, 44274 NANTES CEDEX 2 ; nicolas.sigler@insee.fr*

Résumé. L'objet de cet article est de comparer deux méthodes de formation des classes de repondération utilisées pour corriger la non-réponse totale : la méthode des scores, fondée sur la modélisation de la probabilité de réponse à l'aide de procédures logistiques, et la segmentation. Cette comparaison a été mise en œuvre à partir de simulations sur l'enquête concernant les technologies de l'information et de la communication et le commerce électronique (TIC) 2011. Ont été utilisés 21 scénarii de non-réponse, issus de 7 mécanismes de réponse calés sur 3 taux de réponse (70 %, 80 % et 90 %).

La segmentation semble présenter de nombreux avantages. En termes de formation des classes de repondération, son intérêt essentiel est de fournir des groupes de réponse homogènes ayant un sens du point de vue économique. La méthode, descriptive, permet de les caractériser facilement. Elle est également plus lisible, le nombre de groupes reflétant l'homogénéité des unités en termes de comportement de réponse. En revanche, la qualité des estimateurs obtenus ne donne pas d'avantage net à la segmentation par rapport à la méthode des scores. Les résultats fournis par les deux méthodes sont en effet très proches.

Sommaire

L'objet de cette contribution est de comparer deux méthodes de formation des classes de repondération utilisées pour corriger la non-réponse totale : la méthode des scores, fondée sur la modélisation de la probabilité de réponse à l'aide de procédures logistiques, et la segmentation. Cette comparaison a été mise en œuvre à partir de simulations sur l'enquête concernant les technologies de l'information et de la communication et le commerce électronique (TIC) 2011.

1. Éléments de contexte

1.1 Pourquoi corriger la non-réponse ?

L'absence de réponse, pour tout ou partie d'un questionnaire, a un impact sur les estimateurs produits. La conséquence principale est de **biaiser les estimateurs ponctuels** (proportions pour les variables qualitatives, total ou moyenne par exemple pour les variables quantitatives), si les non-répondants ont un comportement différent des répondants. En effet, dans ce cas, observer les estimateurs sur les seuls répondants ne reflète pas la situation qui aurait été observée sur l'ensemble de l'échantillon. La non-réponse a également pour effet d'**augmenter la variance des estimateurs** : les estimateurs sont moins précis, dans la mesure où ils sont calculés sur un échantillon plus petit, en l'absence de réponse sur une partie de l'échantillon. Enfin, la non-réponse entraîne un **biais des estimateurs de variance standard**, ce qui peut avoir un impact lors de la réalisation de tests par exemple.

1.2 Correction de la non-réponse totale par repondération

Usuellement, pour les enquêtes thématiques entreprises, la non-réponse totale est corrigée par repondération, qui consiste à augmenter le poids de sondage des unités répondantes pour compenser l'absence des unités non répondantes.

Introduisons quelques notations pour expliciter les effets de la non-réponse.

Soit une population U de taille N.

Supposons que l'on veuille estimer le total $Y = \sum_{i \in U} y_i$ pour une variable d'intérêt y.

Pour estimer y, on tire un échantillon aléatoire s, de taille n selon un plan de sondage $p(.)$.

Soit $\hat{Y}_{NR}$ l'estimateur de Y obtenu après correction de la non-réponse totale.

L'erreur totale de $\hat{Y}_{NR}$, $\hat{Y}_{NR} - Y$, peut être décomposée comme la somme de deux termes d'erreur :

$$\hat{Y}_{NR} - Y = \underbrace{\left(\hat{Y}_{HT} - Y\right)}_{\substack{\text{erreur due} \\ \text{à l'échantillonage}}} + \underbrace{\left(\hat{Y}_{NR} - \hat{Y}_{HT}\right)}_{\substack{\text{erreur due} \\ \text{à la non-réponse}}},$$

où $\hat{Y}_{HT}$ est l'estimateur de Horvitz-Thompson que l'on aurait obtenu en l'absence de non-réponse, lequel est un estimateur sans biais de Y sous $p(s)$.

Le biais de l'estimateur s'écrit alors de la manière suivante :

$$\begin{aligned} Biais\left(\hat{Y}_{NR}\right) &= E\left(\hat{Y}_{NR} - Y\right) = E_p E_r\left(\hat{Y}_{NR} - Y \middle| s\right) = E_p\left(\hat{Y}_{HT} - Y \middle| s\right) + E_p E_r\left(\hat{Y}_{NR} - \hat{Y}_{HT} \middle| s\right) \\ &= E_p E_r\left(\hat{Y}_{NR} - \hat{Y}_{HT} \middle| s\right) = E_p\left(B_r\right) \end{aligned}$$

où $B_r = E_r\left(\hat{Y}_{NR} - \hat{Y}_{HT} \middle| s\right)$ est le biais de non-réponse conditionnel étant donné l'échantillon s, $E_p(.)$ l'espérance par rapport au plan de sondage et $E_r(.)$ l'espérance par rapport au mécanisme de réponse.

Le biais de non-réponse conditionnel est nul lorsque le mécanisme de réponse est **uniforme**, c'est-à-dire que pour toutes les unités de la population, la probabilité de réponse est indépendante des variables auxiliaires comme des variables d'intérêt. L'hypothèse d'un mécanisme de réponse uniforme n'est pas réaliste en pratique. Pour s'en approcher, l'idée est de découper l'échantillon en classes, telles que le mécanisme de réponse soit homogène à l'intérieur de chacune d'entre elles. La difficulté est que les probabilités de réponse p_i de chaque unité *i* ne sont pas connues. Dès lors, il faudra modéliser cette probabilité de réponse et utiliser sa valeur estimée pour repondérer au sein de chaque classe.

Si l'échantillon s est divisé en C classes, $s_1,...,s_c,...,s_C$ telles que $s=\bigcup_{c=1}^{C} s_c$, alors le poids ajusté w_i^* pour l'unité i dans la classe c est donné par $w_i^* = w_i / \hat{p}_c$, où $\hat{p}_c$ est le taux de réponse dans la classe c.

L'estimateur par repondération est alors donné par $\hat{Y}_{RC} = \sum_{c=1}^{C} \hat{N}_c \bar{y}_{rc}$,

où $\hat{N}_c = \sum_{i \in s_c} w_i$ et $\bar{y}_{rc} = \sum_{i \in s_c} w_i a_i y_i \Big/ \sum_{i \in s_c} w_i a_i$, soit la moyenne des répondants dans la classe c.

Le biais de non-réponse conditionnel s'écrit sous la forme :

$$Biais\left(\hat{Y}_{RC}|s\right) = E_R\left(\hat{Y}_{RC} - \hat{Y}|s\right) = \sum_{c=1}^{C} \frac{1}{\bar{p}_c} \sum_{i \in s_c} w_i \left(p_i - \bar{p}_c\right)\left(y_i - \bar{y}_c\right),$$

où $\bar{p}_c = \sum_{i \in s_c} w_i p_i \Big/ \sum_{i \in s_c} w_i$ et $\bar{y}_c = \sum_{i \in s_c} w_i y_i \Big/ \sum_{i \in s_c} w_i$.

Ainsi, pour que le biais soit nul, il suffit que le mécanisme de réponse soit uniforme au sein de chaque classe, c'est-à-dire $\hat{p}_i = \hat{p}_c$ si l'unité $i \in s_c$.

Dans la pratique, lorsque les classes - ou « groupes de réponse homogènes » (GRH) - sont constituées, il reste alors à repondérer les observations. On calcule tout d'abord le taux de réponse observé à l'intérieur des classes, comme le rapport entre le nombre de répondants et l'effectif total du groupe (soit ici le nombre d'entreprises en tenant compte de leur pondération déterminée à l'échantillonnage). Le taux de réponse pour la classe c est ainsi donné par :

$$Taux\ de\ réponse = \frac{\sum_{répondantes} w_i}{\sum_{ensemble\ de\ l'échantillon} w_i} = \hat{p}_c .$$

Enfin, les poids déterminés suite à la correction de la non-réponse sont calculés à l'intérieur de chaque classe en divisant le poids de départ des répondants par ce taux de réponse observé (estimation de la probabilité de réponse) : $w_i^* = w_i / \hat{p}_c$.

Notons que cette étape de repondération est menée de manière identique, quelle que soit la méthode de formation des classes utilisée.

1.3 Les méthodes de formation des classes de repondération

Différentes méthodes peuvent être mobilisées pour former les classes homogènes (ou GRH) par rapport aux probabilités de réponse. Elles sont fondées sur de la modélisation ou de la segmentation, mais reposent toutes sur la sélection au préalable d'un ensemble de variables auxiliaires (disponibles pour toutes les unités de l'échantillon, qu'elles soient répondantes ou non répondantes) liées au comportement de réponse.

La mise en œuvre de ces méthodes peut être résumée de la manière suivante :

- Techniques de modélisation :
    - Méthode des croisements :

Cette méthode consiste à modéliser les probabilités de réponse p_i à partir d'un modèle contenant au départ toutes les interactions possibles entre les variables auxiliaires préalablement sélectionnées et catégorisées si besoin. Les croisements de variables sont ensuite regroupés de manière itérative jusqu'à former les classes de repondération. Des contraintes sont imposées sur le nombre et la proportion d'unités répondantes dans chaque classe, afin d'éviter les classes à faible effectif, qui pourraient rendre les estimateurs instables.

La méthode des croisements est celle qui s'approche le plus de celle mise en œuvre jusqu'à présent au pôle Ingénierie statistique Entreprises de l'Insee. Elle nécessite néanmoins une étape d'expertise dans la définition du modèle, afin d'obtenir des groupes de réponse ayant un sens d'un point de vue économique, ou tout au moins qui soient interprétables selon différentes variables. Pour cette raison, nous avons pour nos simulations choisi d'utiliser une méthode alternative, plus facilement automatisable : la méthode des scores.

    - Méthode des scores :

La méthode des scores s'appuie également sur de la modélisation, mais sans introduire les interactions entre les variables auxiliaires. La probabilité de réponse est estimée pour toutes les unités de l'échantillon, répondantes ou non, et sert de critère d'homogénéité pour la formation des classes. Soit les estimateurs $\hat{p}_i$ sont ordonnés pour diviser l'échantillon en classes de tailles égales (méthode des « quantiles égaux »), soit les unités similaires sont regroupées à l'aide d'une classification.

- Techniques de segmentation :

Dans l'approche par segmentation, la population de départ est découpée de manière successive selon les modalités des variables déterminées comme les plus discriminantes à chaque itération. Différents algorithmes de segmentation existent, tel que l'algorithme CHAID (Kass, 1980).

1.4 Comparaison des deux méthodes

Notre objectif est ici de comparer les deux méthodes de correction de la non-réponse : la méthode traditionnelle, par régression logistique, et la méthode par segmentation. La comparaison est effectuée à partir de simulations sur l'enquête concernant les technologies de l'information et de la communication et le commerce électronique (enquête TIC, cf. annexe 4), en combinant différents scénarii de non-réponse sur 1 000 échantillons issus de l'enquête 2011.

Le protocole suivi se déroulera suivant les différentes étapes décrites ci-dessous :

- constitution d'une population fictive à partir de l'enquête TIC 2011
- calcul des estimateurs cibles
- échantillonnage
- génération de la non-réponse
- correction de la non-réponse selon les deux méthodes
- analyse des estimateurs obtenus (biais, variance)

Les estimateurs sont comparés au regard du biais relatif $\left(RB_{MC}\right)$ et de l'erreur quadratique moyenne $\left(MSE_{MC}\right)$ de Monte-Carlo.

On compare ainsi :

$$RB_{MC}\left(\hat{Y}^{seg}\right)=\frac{1}{R}\sum_{j=1}^{R}\frac{\hat{Y}_j^{seg}-Y}{Y}\times 100 \text{ (en \%)} \quad \text{et} \quad RB_{MC}\left(\hat{Y}^{score}\right)=\frac{1}{R}\sum_{j=1}^{R}\frac{\hat{Y}_j^{score}-Y}{Y}\times 100 \text{ (en \%)},$$

$$MSE_{MC}\left(\hat{Y}^{seg}\right)=\frac{1}{R}\sum_{j=1}^{R}\left(\hat{Y}_j^{seg}-Y\right)^2 \quad \text{et} \quad MSE_{MC}\left(\hat{Y}^{score}\right)=\frac{1}{R}\sum_{j=1}^{R}\left(\hat{Y}_j^{score}-Y\right)^2$$

(on compare en fait $\frac{MSE_{MC}\left(\hat{Y}^{seg}\right)}{MSE_{MC}\left(\hat{Y}^{score}\right)}$ à 1)

où R est le nombre d'échantillons (1 000 dans notre cas), $\hat{Y}_j^{seg}$ et $\hat{Y}_j^{score}$ les estimateurs des méthodes *segmentation* et *scores* calculés sur l'échantillon j, et Y l'estimateur cible.

2. Étapes préliminaires

2.1 Situation de référence

Avant de procéder aux simulations, comprenant une phase d'échantillonnage et la génération de la non-réponse, la première étape consiste à créer une population fictive à partir du fichier TIC 2011 de fin d'enquête, dans lequel les non-réponses partielle et totale ont été corrigées.

2.1.1 La population fictive

La population fictive est créée à partir de l'échantillon de TIC 2011. Les entreprises conservées pour créer cette population artificielle, après redressement de la non-réponse totale et partielle sur le fichier d'origine, sont les entreprises répondantes ou considérées comme telles. Ce dernier cas concerne les entreprises **non substituables** : ce sont de très grandes entreprises et/ou des entreprises très particulières par rapport aux thèmes traités, pour lesquelles des traitements particuliers sont effectués. Au total, 10 062 entreprises sont retenues, qui représentent 81 % de l'échantillon d'origine. Ces unités sont dupliquées selon leur poids de calage, pour obtenir une population de départ proche de la population d'origine. Au final, la population fictive est composée de 182 461 entreprises.

2.1.2 Les estimateurs cibles

Pour comparer les deux méthodes, ont été choisies 14 variables test : 5 variables quantitatives et 9 variables qualitatives (oui/non). Les estimateurs cibles calculés correspondent au total pour les variables quantitatives et à la proportion de « oui » pour les variables qualitatives.

Variables test et estimateurs cibles :

		Cible
Variables numériques		
A2	Nombre de personnes utilisant un ordinateur	5 890 389
B4	Nombre de personnes utilisant Internet	4 679 980
G2	Montant du CA généré via le web (en milliers d'euros)	70 092 907
G5	Montant du CA généré via EDI* (en milliers d'euros)	304 706 250
G8	Montant des achats électroniques (en milliers d'euros)	266 397 118
Variables qualitatives (proportion de 'oui', en %)		
B1	Présence d'un accès Internet	98,3
B6	Présence d'un site web ou d'une page d'accueil	62,7
C1	Présence d'un système d'échange électronique traité automatiquement	48,1
D1a	Factures électroniques aux clients par traitement automatique	9,6
D1b	Factures électroniques aux clients par courrier ou pièces jointes pdf	34,8
G1	Réception de commandes de biens ou services sur le site web	9,3
G4	Réception de commandes de biens ou services via EDI*	5,3
G7	Achat de biens et services par voie électronique	27,4
H1	Utilisation d'outils fondés sur la RFID*	2,5

* EDI = Echange de données informatisées, RFID = Identification par radio-fréquence

2.2 Génération des échantillons

Avant de simuler la non-réponse, nous procédons à une phase d'échantillonnage, l'idée étant de cumuler les mécanismes aléatoires comme en pratique dans les enquêtes. Après avoir créé la population fictive, la seconde étape consiste à tirer 1 000 échantillons à partir de cette population servant de base de sondage, le nombre d'échantillons étant suffisamment important pour que les différences observées par la suite ne soient pas le fait du hasard. On reproduit l'échantillonnage stratifié par secteur et taille de l'enquête TIC 2011. Les entreprises d'au moins 500 personnes occupées sont interrogées exhaustivement. Pour les autre strates, les taux de sondage proviennent d'une allocation proportionnelle au nombre de personnes occupées.

Cette étape a été réalisée à l'aide de la procédure Surveyselect de SAS, en entrant le nombre d'unités à tirer dans chaque strate. Chacun des 1 000 échantillons contient *in fine* 12 028 entreprises.

2.3 Génération de la non-réponse

Après avoir réalisé l'échantillonnage, l'étape suivante consiste à générer la non-réponse. Nous avons simulé 21 scénarii de non-réponse, générés à partir de 7 mécanismes de réponse calés sur 3 taux de réponse différents (70 %, 80 % et 90 %).

Les mécanismes de réponse ont été construits selon différentes logiques. Nous avons eu recours au secteur et à la taille des unités, deux variables clés des enquêtes thématiques entreprise. Nous avons également testé un mécanisme fondé sur les groupes de réponse homogènes, tels qu'ils ont été déterminés lors du traitement effectif de l'enquête TIC 2011. Nous avons par ailleurs souhaité tester des mécanismes plus proches de la réalité en introduisant des « variables cachées », au sens où celles-ci ne seront pas utilisées ensuite pour corriger la non-réponse.

Les 7 mécanismes de réponse sont les suivants :

- **Aléatoire simple sans remise**
- **ZAU** : localisation de l'unité selon la typologie des communes en aires urbaines ; il s'agit d'une variable cachée
- **Taux d'endettement** : taux d'endettement de l'entreprise catégorisé en déciles ; il s'agit d'une variable cachée
- **GRH** : les GRH retenus lors du redressement de l'enquête TIC 2011, au nombre de 18 ; les variables composantes sont la région (DOM/TOM, Paris et petite couronne, province), le comportement de réponse à l'enquête précédente, le secteur (10 postes) et la taille (5 tranches)
- **GRH x taux d'endettement** : croisement des variables précédentes (taux d'endettement en variable cachée)
- **Secteur x taille** : croisement du secteur (regroupé en 23 postes de la nomenclature agrégée A38) et de la taille en 5 tranches (10 à 19, 20 à 49, 50 à 249, 250 à 499, 500 salariés et plus)
- **Secteur x taille x taux d'endettement** : croisement des variables précédentes (taux d'endettement en variable cachée).

Ces mécanismes de réponse ont été générés en fixant une probabilité de réponse pour chaque modalité des variables (ou croisement de variables) considérées, en reproduisant le taux de réponse observé sur l'enquête TIC 2011. Cette première phase a été réalisée sous SAS à l'aide de la procédure Surveyselect. La macro Calmar a ensuite été utilisée pour caler les taux de répondants à 70 %, 80 % et 90 %.

3. Première méthode : la méthode des scores

La première méthode de correction proposée, la méthode des scores, est mise en œuvre avec la méthode des quantiles égaux, qui permet d'automatiser les traitements et d'être plus « objectif », par rapport à la pratique usuelle dans le traitement des enquêtes, où une expertise est menée au cas par cas pour amender les classes de repondération.

3.1 Description

La méthode des scores, permettant de construire les classes de repondération, procède suivant les deux phases décrites ci-dessous.

Étape 1 : Modélisation du score

La 1ère étape consiste à modéliser la probabilité de réponse p_i.

Rappelons que nous sommes dans la situation d'un mécanisme de réponse défini par une indicatrice de réponse a_i qui vaut 1 si l'individu i répond et 0 sinon, telle que : $a_i = \begin{cases} 1 & avec\ une\ probabilité\ p_i \\ 0 & avec\ une\ probabilité\ 1-p_i \end{cases}$.

À partir de variables auxiliaires disponibles pour toutes les unités dans l'échantillon (répondantes et non répondantes), on estime la probabilité de réponse p_i pour toutes les unités, qu'elles soient répondantes ou non répondantes, à l'aide d'un modèle logistique de la forme :

$\log\left(\frac{p_i}{1-p_i}\right) = z_i'\beta$, où z_i est un vecteur de variables auxiliaires disponibles pour toutes les unités dans l'échantillon et β un vecteur de paramètres inconnus.

On obtient alors la prédiction $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = z_i'\hat{\beta}$ pour toutes les unités dans l'échantillon, le vecteur $\hat{\beta}$ étant obtenu par la méthode du maximum de vraisemblance.

On en déduit le score $\hat{p}_i$, qui servira de critère d'homogénéité des classes. Notons que lors de cette étape, est également déterminé l'ensemble des variables auxiliaires qui expliquent le comportement de réponse.

Étape 2 : Formation des classes

En se fondant sur les probabilités de réponse estimées $\hat{p}_i$, on forme les classes suivant la méthode des quantiles égaux, en répartissant les unités en un certain nombre de groupes, défini *a priori*. Pour ce faire, on ordonne les valeurs de $\hat{p}_i$ en ordre croissant, puis on divise l'échantillon en classes de tailles approximativement égales.

L'idée est que si les probabilités de réponse p_i sont bien estimées, alors $\hat{p}_i \approx p_i$, et les classes, homogènes par rapport à $\hat{p}_i$, le seront également par rapport aux p_i, de sorte que le biais sera proche de 0.

3.2 Intérêt de la méthode

Cette méthode, automatique, présente des avantages :

- L'information des variables auxiliaires est résumée dans l'estimation de la probabilité p_i ; dès lors, le problème de la présélection de variables auxiliaires pertinentes ne se pose pas.
- La modélisation par groupes amène une certaine robustesse si le modèle est mal spécifié.
- On maîtrise *a priori* le nombre de classes, ce qui permet d'assurer d'un nombre suffisant d'individus par classe, même si, en théorie, le choix du nombre de classes peut s'avérer

délicat. En effet, il résulte d'un compromis entre, d'une part, augmenter le nombre de classes pour réduire le biais en assurant une plus grande homogénéité à l'intérieur des classes, et, d'autre part, diminuer le nombre de classes pour avoir davantage de répondants dans chacune et donc une meilleure précision pour des estimateurs plus stables.

Les inconvénients sont les suivants :

- Le principal est que la répartition par quantiles égaux peut amener à regrouper des unités très différentes (secteur, taille, etc.). Aussi, les classes peuvent ne pas avoir de cohérence économique, ce qui peut, en pratique, dérouter quelque peu les maîtrises d'ouvrage des enquêtes.
- En théorie, il faudrait également accorder une importance particulière, en amont de la procédure Logistic, à la structure des données en termes de découpage et de regroupement des modalités. Cette phase préalable n'a toutefois pas été réalisée ici, de manière à comparer les deux méthodes de correction de la non-réponse à partir de variables identiques en entrée.
- Il faut également signaler que la méthode ne permet pas de détecter les interactions entre variables, sauf à les intégrer explicitement en entrée de la modélisation.

3.3 Mise en œuvre

La modélisation des indicatrices de réponse et l'estimation des probabilités de réponse sont réalisées sous SAS à l'aide d'une procédure Logistic non pondérée. Les variables en entrée du modèle sont des variables qualitatives disponibles pour toutes les unités (souvent des variables de lancement de l'enquête) ayant potentiellement une influence sur la probabilité de réponse. Le choix a été largement inspiré de la constitution des groupes de réponse homogènes réalisée lors du traitement effectif de la non-réponse sur le fichier TIC 2011, qui retenait les variables : secteur d'activité, taille, localisation, comportement de réponse à l'enquête précédente et appartenance à un groupe. Nous avons ajouté le chiffre d'affaires, et veillé à prendre le même niveau de détail que lors de la génération de la non-réponse pour les variables relatives au secteur et à la taille de l'entreprise.

Au final, les variables proposées en entrée de la modélisation sont les suivantes :

- le secteur d'activité :
  - regroupement en 23 postes de la nomenclature d'activité agrégée A38
- la taille de l'entreprise :
  - 10 à 19 salariés
  - 20 à 49 salariés
  - 50 à 249 salariés
  - 250 à 499 salariés
  - 500 salariés et plus
- la localisation géographique :
  - DOM/TOM
  - Paris et petite couronne
  - autres régions regroupées par grandes zones géographiques
- le comportement de réponse à l'enquête précédente :
  - entreprise non interrogée dans TIC 2010
  - entreprise interrogée répondante à TIC 2010
  - entreprise interrogée non répondante ou hors champ dans TIC 2010
- l'appartenance à un groupe en 2008 :
  - oui
  - non

- le chiffre d'affaires (en milliers d'euros) :
  - 0 à moins de 2 000
  - 2 000 à moins de 5 000
  - 5 000 à moins de 20 000
  - 20 000 à moins de 70 000
  - 70 000 et plus.

Les groupes de réponse homogènes sont ensuite déterminés à l'aide de la procédure Rank de SAS, le nombre de classes étant paramétré à 25. Il faut préciser que le nombre de classes demandées constitue en fait un maximum. En pratique, le nombre de classes peut être inférieur à 25 en fonction du modèle trouvé lors de la procédure Logistic (cf. page 15).

4. Deuxième méthode : la segmentation par arbre

4.1 Présentation de la segmentation par arbre

La segmentation consiste à construire des groupes d'unités les plus homogènes possible par rapport à une variable d'intérêt Y en utilisant l'information de p variables $X_1, ..., X_p$, dites « explicatives ».

Elle procède par divisions successives des unités d'une population en segments, ou « nœuds », construisant un arbre, de sorte que chaque nœud soit homogène par rapport à la variable d'intérêt en utilisant l'information des variables explicatives. L'ensemble des nœuds terminaux, ou « feuilles », constitue une partition de la population initiale en classes homogènes par rapport à la variable d'intérêt.

Les variables Y et $X_1, ..., X_p$ peuvent être binaires, nominales, ordinales ou quantitatives. Si Y est qualitative (binaire, nominale ou ordinale), on parle d'arbre de classification. Si Y est quantitative, on parle d'arbre de régression.

Diverses méthodes de segmentation par arbres (CART, ID3, C4.5 et C5.0, CHAID, QUEST, etc.) ont été proposées depuis les années 1960. Elles diffèrent par le type de variables à exploiter (qualitatives, continues…), par l'indicateur de qualité ou les critères d'arrêt retenus.

Ces méthodes sont disponibles dans différents logiciels de statistique. Pour automatiser la segmentation sur 1 000 échantillons pour chacun des 21 scénarii, nous avons choisi d'utiliser la méthode CHAID, implémentée en SAS via la macro TREEDISC. Celle-ci permet de générer un code SAS utilisable pour réaliser une classification des observations.

La méthode CHAID (Chi-square Automatic Interaction Detection) a été proposée par Kass (1980). Il s'agit d'une amélioration des algorithmes AID (Morgan et Sonquist, 1963) et THAID (Messenger et Morgan, 1973). La méthode fonctionne avec des variables qualitatives ou quantitatives, et repose sur l'utilisation de deux algorithmes : pour regrouper les modalités des variables explicatives et pour construire l'arbre.

Nous expliquons ci-dessous le processus suivi avec des variables qualitatives.

Soient Y la variable d'intérêt qualitative à K modalités et $X_1, ..., X_J$, J variables explicatives qualitatives, telles que X_j possède M_j modalités.

Étape 1 : Algorithme de regroupement des modalités

L'algorithme est fondé sur des tests du chi2, dont la statistique se calcule de la manière suivante, pour deux variables qualitatives déclinées respectivement selon p et q modalités :

$$\chi^2 = \sum_{k=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{kj} - \frac{n_{k.} \times n_{.j}}{n}\right)^2}{\frac{n_{k.} \times n_{.j}}{n}}$$

On construit pour chaque variable explicative X_j le tableau de contingence de dimension (K, M_j) croisant Y et X_j :

Y/X_j	X_j^1 ... x_j^m ... x_j^{Mj}	Σ
y_1		
…		
y_k		
…		
y_K		
Σ		n_j

On détermine la paire de modalités de X_j dont le sous-tableau $(K,2)$ fournit le plus petit chi2.Si ce chi2 n'est pas significatif, c'est-à-dire que la liaison entre Y et les deux modalités considérées de X_j est faible, on regroupe les deux modalités correspondantes en une seule modalité et on répète l'opération jusqu'à ce que toutes les modalités (regroupées ou non) présentent un chi2 significatif.

Ensuite, pour chaque modalité composée de plus de trois modalités originales, on détermine la division binaire associée au chi2 le plus grand. Si ce chi2 est significatif, c'est-à-dire que la liaison entre Y et les deux nouvelles modalités est forte, on effectue cette division binaire et on revient au début de l'étape 1.

Étape 2 : Algorithme de construction de l'arbre

Il faut tout d'abord trouver la variable explicative la plus significative au moyen de tests du chi2. Pour cela, on calcule la significativité p' de chaque variable X_j dont les modalités ont été regroupées et on retient la plus significative.

Le degré de significativité corrigé p' est obtenu en multipliant le degré p du test du chi2 du tableau réduit par le coefficient de Bonferroni, qui représente le nombre de possibilités de regrouper les L modalités d'une variable explicative en g groupes $(1 \le g \le L)$ et vaut selon le type de variable :

- nominale : $B_{nom} = \sum_{i=0}^{g-1} (-1)^i \frac{(g-i)^L}{i!(g-i)!}$,
- ordinale : $B_{ord} = \binom{L-1}{g-1}$,
- ordinale avec une modalité « flottante » (sans objet) : $B_{ord} = \binom{L-2}{g-2} + g\binom{L-2}{g-1}$.

Si la significativité dépasse la valeur seuil définie *a priori*, on divise l'ensemble des observations en autant de segments que de modalités composées de la variable choisie.

Pour chaque segment ainsi obtenu, on répète ensuite l'étape 2 jusqu'à ce qu'il n'y ait plus de variable explicative significative.

4.2 Intérêt de la méthode

La segmentation par arbre présente de nombreux avantages :

- Le principal avantage est la simplicité de son fonctionnement, par divisions successives de la population. La lisibilité des règles d'affectation des unités aux groupes permet de communiquer aisément sur la méthode, en l'illustrant notamment par des arbres.
- La segmentation est une méthode non paramétrique, ne nécessitant pas d'hypothèses sur la distribution des variables.
- Elle est peu sensible aux valeurs extrêmes ou aberrantes.
- Elle permet de détecter les interactions entre plusieurs variables.
- Par ailleurs, comme pour la méthode des scores, le problème de la présélection de variables explicatives pertinentes ne se pose pas, puisque la segmentation met en œuvre des tests de significativité des variables.
- Par rapport à la méthode des scores, la segmentation devrait permettre d'obtenir des groupes ayant davantage de cohérence d'un point de vue économique, dans la mesure où les modalités des variables qualitatives ordinales sont regroupées de manière adjacente, tandis qu'avec la méthode des scores, tous les regroupements sont possibles.

La segmentation n'est toutefois pas exempte d'inconvénients :

- Le principal, tout au moins avec la macro TREEDISC, est que l'on ne maîtrise pas le nombre de groupes de réponse homogènes constitués, qui peuvent en outre être de taille très différente.
- Il s'agit d'une méthode « descriptive » et non « explicative », ce qui signifie qu'elle ne s'appuie sur aucun test, puisqu'aucun modèle n'est supposé *a priori*.
- Par ailleurs, la méthode est assez sensible à la structure des données (nombre de modalités des variables, etc.).
- Elle peut également manquer de robustesse, si l'échantillon n'est pas assez grand.

4.3 Mise en œuvre

L'algorithme de la macro TREEDISC sous SAS est similaire à l'algorithme CHAID décrit plus haut, même si les critères de significativité utilisés lors des tests sont différents : pour le regroupement des modalités (étape 1), la significativité des tests est observée au regard des p-values ajustées. Par ailleurs, pour trouver la variable de division la plus significative (étape 2), l'algorithme propose une amélioration de l'ajustement de Bonferroni.

La macro TREEDISC permet de réaliser des arbres m-aires. Les variables explicatives peuvent être qualitatives (nominales ou ordinales, avec possibilité de modalité « flottante » (sans objet)) ou quantitatives. Lors du partitionnement, les modalités des variables nominales peuvent être regroupées sans contrainte. Pour les variables ordinales, les modalités, étant ordonnées, ne peuvent être regroupées qu'entre modalités adjacentes. Les valeurs manquantes sont traitées comme une modalité supplémentaire.

Les différents critères (macro variables) à renseigner sont :

- le seuil de significativité du test du chi2 (alpha)

- les critères d'arrêt :
    - le nombre minimum d'observations d'un nœud pour qu'il soit subdivisé en plusieurs branches (branch)
    - le nombre minimum d'observations pour constituer une nouvelle feuille (leaf)
    - le nombre maximum de niveaux de l'arbre (maxdepth).

Différents tests ont montré que pour notre jeu de données, il était difficile de concilier un seuil de significativité faible (moins de 10 %) avec un nombre relativement important de groupes de réponse homogènes (au moins une vingtaine), sans que ces derniers soient trop volumineux (moins de 2 500 unités). Au final, les segmentations, non pondérées, ont été réalisées avec le seuil de significativité alpha=0,1 et les critères d'arrêt suivants : branch=160, leaf=80 et maxdepth=20.

5. Analyse des résultats

5.1 Analyse des « modèles » trouvés

5.1.1 Les variables retenues dans le modèle

Pour comparer les performances des deux méthodes, nous nous intéressons dans un premier temps aux « modèles » trouvés via la régression logistique et la segmentation. Le terme de « modèle » est ici un abus de langage, puisque dans la seconde méthode, il s'agit plus exactement des variables intervenant dans les segmentations, aucune modélisation n'étant réalisée.

Globalement, quelle que soit la méthode utilisée, les résultats montrent une différence notable entre les scénarii 'aléatoire simple' et 'variable cachée seule' (ZAU ou taux d'endettement), où le plus souvent aucun modèle n'est trouvé, sauf avec un taux de réponse de 90 %, et les autres scénarii, pour lesquels un modèle est trouvé dans tous les cas. Les mécanismes avec variables cachées seules sont approchés par des modèles comprenant une seule variable : le secteur d'activité prédomine pour le taux d'endettement et la localisation géographique pour le ZAU, mais le nombre d'échantillons concernés reste assez modeste pour le taux d'endettement. Pour les autres scénarii, les modèles sont en général bien reconstitués ou approchés, même s'ils sont combinés à une variable cachée. Dans ce cas, les variables cachées renforcent l'effet des variables avec lesquelles elles sont les plus corrélées. Par ailleurs, pour tous les scénarii, plus le taux de réponse augmente, plus les modèles auront tendance à comporter davantage de variables. Ainsi, des taux de réponse élevés n'assurent pas de retrouver exactement le mécanisme de réponse, par contre, ils permettent de repérer davantage de modèles cohérents avec celui-ci.

Les modèles observés avec la régression logistique et la segmentation sont très proches : globalement, les méthodes explicative et descriptive se rejoignent. La principale différence est que la segmentation garde davantage de variables par rapport à la modélisation logistique, en particulier dans les scénarii avec 'GRH' et quand le taux de réponse est élevé. Les modèles trouvés avec la segmentation sont donc beaucoup plus variés et comportent souvent davantage de variables, au point que pour les scénarii fondés sur le GRH, avec ou sans la variable cachée, les deux tiers des modèles avec 90 % de réponse comprennent l'ensemble des variables potentielles. La plus grande variété de modèles avec la segmentation est également de mise pour les scénarii reposant sur le secteur et la taille de l'entreprise, qu'ils soient associés ou non au taux d'endettement.

Principaux modèles trouvés par régression logistique selon le scénario (en % sur les 1 000 échantillons) :

Modélisation			Modèles trouvés : variables significatives													Autres modèles	Ensemble des modèles
		Secteur			x	x	x	x	x	x	x	x	x	x	x		
		Taille								x	x	x	x	x			
		Localisation géographique		x			x	x	x			x	x	x			
		Comportement de réponse en 2010					x	x	x			x	x	x			
		Appartenance à un groupe							x					x			
		Chiffre d'affaires				x		x			x		x				
Scénario	Aléatoire simple	70 %	75													25	100
		80 %	75													25	100
		90 %	76													24	100
	ZAU	70 %	56	16												28	100
		80 %	37	29												34	100
		90 %		53												47	100
	Taux d'endettement	70 %	65													35	100
		80 %	60													40	100
		90 %	41		19											40	100
	GRH	70 %					41	20	22			12				5	100
		80 %					15	28	25			26				6	100
		90 %						27				25	18	11		19	100
	GRH x taux d'endettement	70 %					41	22	20			11				7	100
		80 %					16	35	28			17				5	100
		90 %						44	11			14	17			15	100
	Secteur x taille	70 %				48				36						16	100
		80 %								74						26	100
		90 %								58	26					16	100
	Secteur x taille x taux d'endettement	70 %				64				22						14	100
		80 %				44				41						15	100
		90 %				41				37						22	100

Note : Les cases blanches représentent moins de 10 % des échantillons.

Lecture : Lorsque le mécanisme de réponse est fondé sur les variables secteur x taille avec un taux de réponse de 70 %, pour 48 % des échantillons, le modèle trouvé comprend le secteur et le chiffre d'affaires.

Principaux « modèles » issus de la segmentation selon le scénario (en % sur les 1 000 échantillons) :

Variables			"Modèles" trouvés : variables significatives														Autres	Ensemble
		Secteur				x	x	x	x	x	x	x	x	x	x	x		
		Taille								x	x	x	x	x	x	x		
		Localisation géographique			x			x	x				x	x	x	x		
		Comportement de réponse en 2010						x	x			x		x	x	x		
		Appartenance à un groupe					x		x		x	x	x		x	x		
		Chiffre d'affaires		x		x	x	x	x	x	x	x	x	x		x		
Scénario	Aléatoire simple	70 %	63														37	100
		80 %	65														35	100
		90 %	65														35	100
	ZAU	70 %	44														56	100
		80 %	27		14												59	100
		90 %			19												81	100
	Taux d'endettement	70 %	52														48	100
		80 %	48														52	100
		90 %	29	10													61	100
	GRH	70 %							20					13	15	22	30	100
		80 %							12					12	24	40	12	100
		90 %							11						21	66	2	100
	GRH x taux d'endettement	70 %						12	17					11	13	21	26	100
		80 %							16					15	17	40	12	100
		90 %													13	68	19	100
	Secteur x taille	70 %				13				12							75	100
		80 %									17	16	13			11	45	100
		90 %									19	16	12			13	40	100
	Secteur x taille x taux d'endettement	70 %				19	13										68	100
		80 %								16	18	12					54	100
		90 %									19	19	16			14	32	100

Note : Les cases blanches représentent moins de 10 % des échantillons.

Lecture : Lorsque le mécanisme de réponse est fondé sur les variables secteur x taille avec un taux de réponse de 70 %, pour 13 % des échantillons, le « modèle » trouvé comprend le secteur et le chiffre d'affaires.

5.1.2 Le nombre de groupes de réponse homogènes

Un autre point à examiner, pour mieux comprendre le fonctionnement de la repondération, est le nombre de groupes de réponse homogènes obtenus *in fine*, après la modélisation.

Pour la *méthode des scores*, le nombre de groupes de réponse homogènes est compris entre 1 et 25 selon les scénarii. Le nombre de classes est en effet fixé au maximum à 25, mais varie en fonction du modèle trouvé. Ainsi, reflet des conclusions précédentes, dans le cas du scénario aléatoire simple ou des variables cachées seules, les échantillons constituent souvent une seule classe de repondération. C'est le cas lorsque le modèle est réduit à la constante. Lorsqu'une ou plusieurs variables sont jugées significatives, il est possible d'atteindre 25 groupes de réponse homogènes. Pour les autres scénarii, le nombre minimum de groupes s'établit à 18 (secteur x taille), suffisant pour assurer une bonne homogénéité au sein des classes. On peut remarquer également que plus le mécanisme fait intervenir de nombreuses variables, plus le nombre de groupes est élevé, ce qui est le cas lorsqu'on ajoute la variable cachée.

Avec la *segmentation*, le nombre de groupes de réponse varie entre 1 et 35. Contrairement à la méthode des scores, le nombre de classes maximum n'est ici pas imposé. Cela a peu d'impact sur le scénario aléatoire simple et sur ceux fondés sur les variables cachées seules, où le nombre de groupes est relativement faible, comme avec la méthode des scores. Par contre, pour les autres scénarii, la segmentation aboutit à un nombre de groupes de réponse homogènes très variable, même pour un scénario donné. Ainsi, pour les scénarii fondés sur le GRH ou le croisement secteur x taille, avec ou sans variable cachée, tandis que pour la méthode des scores le nombre de classes est compris entre 18 et 25, avec la segmentation, celui-ci oscille entre 3 et 35. Si le nombre de GRH défini par la méthode des scores dépend principalement de son paramétrage, le fait que le nombre de GRH déterminés par la segmentation ne soit pas contraint permet une meilleure adéquation entre ce nombre de GRH et la situation à décrire.

Nombre de GRH obtenus par scénario selon la méthode :

Scénario		Méthode des scores			Segmentation		
Mécanisme de réponse	**Taux de réponse**	**Minimum**	**Médiane**	**Maximum**	**Minimum**	**Médiane**	**Maximum**
Aléatoire simple	70 %	1	1	25	1	1	16
	80 %	1	1	25	1	1	12
	90 %	1	1	25	1	1	12
ZAU	70 %	1	1	25	1	2	14
	80 %	1	5	25	1	3	15
	90 %	1	9	25	1	4	17
Taux d'endettement	70 %	1	1	25	1	1	13
	80 %	1	1	25	1	2	17
	90 %	1	3	25	1	3	15
GRH	70 %	21	25	25	6	11	22
	80 %	25	25	25	7	13	24
	90 %	20	23	25	10	16	30
GRH x taux d'endettement	70 %	21	25	25	5	11	22
	80 %	25	25	25	7	13	23
	90 %	23	25	25	9	17	29
Secteur x taille	70 %	19	25	25	3	8	19
	80 %	18	24	25	6	14	24
	90 %	18	24	25	16	24	35
Secteur x taille x taux d'endettement	70 %	20	25	25	3	8	17
	80 %	23	25	25	6	13	28
	90 %	23	25	25	9	17	30

5.2 Analyse des estimateurs

La qualité des estimateurs obtenus après correction de la non-réponse est analysée au regard des indicateurs de Monte-Carlo (cf. pages 5 et 6).

Concernant les *variables quantitatives*, qui sont les principales variables sur lesquelles les résultats de l'enquête sont diffusés, les estimateurs sont globalement bien redressés par les deux méthodes. Le biais relatif est toujours faible, en valeur absolue inférieur à respectivement 1,3 % et 1,1 % pour la méthode des scores et la segmentation. L'erreur quadratique moyenne est également du même ordre de grandeur par les deux méthodes, avec un rapport compris entre 0,66 et 1,29. L'erreur relative est d'ailleurs particulièrement faible pour les variables de montant de chiffre d'affaires et d'achats électroniques (variables G2, G5 et G8), où elle varie entre 0,93 et 1,08. En termes de scénario, les mécanismes fondés sur le GRH semblent mieux corrigés par la méthode des scores, tandis que les taux de réponse plus élevés (90 %) seraient à l'avantage de la segmentation, même si les écarts restent faibles.

Pour ce qui est des *variables qualitatives*, les résultats sont également très proches via les deux méthodes. Comme pour les variables quantitatives, les biais relatifs sont faibles : ils sont contenus à moins de 1,2 % en valeur absolue, sauf dans le cas du scénario 'secteur x taille x taux d'endettement' pour la réception de commandes de biens ou services sur le site web (G1), où ils atteignent 1,5 % pour la méthode des scores et 2,1 % par la segmentation. Par ailleurs, l'erreur quadratique moyenne est très proche via les deux méthodes : l'erreur relative est comprise entre 0,95 et 1,17, les valeurs extrêmes étant observées pour le scénario 'secteur x taille x taux d'endettement'.

6. Comparaison des méthodes

6.1 La « modélisation »

Les résultats présentés sur la « modélisation » et le nombre de groupes de réponse homogènes constitués traduisent le fonctionnement très différent des deux méthodes. Malgré quelques inconvénients, **la phase de « modélisation » semble être à l'avantage de la segmentation**.

Cette méthode permet en effet de **décrire et interpréter les groupes de réponse homogènes** à l'aide des caractéristiques des unités et de leur taux de réponse. Les règles d'affectation précises rendent la méthode transparente pour l'utilisateur et permettent de communiquer facilement sur les choix réalisés en termes de correction de la non-réponse. En comparaison, avec la méthode des scores, les classes de repondération peuvent regrouper des unités très différentes, puisque ces dernières sont classées uniquement en fonction de leur probabilité prédite. Il faut toutefois préciser que ce constat est lié en partie à la répartition des entreprises par quantiles égaux. Dans le traitement actuel des enquêtes thématiques entreprises, les résultats de la régression logistique sont utilisés pour constituer, « à la main », des groupes d'entreprises présentant des caractéristiques proches selon leur probabilité de réponse. Cela revient à faire pas à pas ce que la segmentation réalise de manière automatique, à ceci près que la phase de modélisation en amont de la méthode des scores risque de rejeter un certain nombre de variables candidates utilisées dans la segmentation.

Il est également intéressant de souligner que la segmentation prend en compte un nombre plus important de variables pour un nombre de groupes de réponse homogènes souvent plus faible qu'avec la méthode des scores. La segmentation permet donc d'**utiliser davantage d'information auxiliaire**. Quant au nombre plus important de classes déterminées avec la méthode des scores, s'il peut donner l'impression de mieux corriger la non-réponse, la répartition par quantiles égaux n'assure aucunement que les probabilités soient très différentes entre les groupes obtenus. Dès lors, le nombre de classes issu de la segmentation a plus de sens que celui obtenu par la méthode des scores, qui peut s'avérer artificiellement élevé. La segmentation est donc également plus transparente sur ce point.

Pour compléter la comparaison, deux aspects négatifs de la segmentation doivent être signalés. Cette méthode nécessite en amont de nombreux tests sur les paramètres à renseigner (profondeur de l'arbre, nombre d'observations minimum par nœud, etc.), et surtout une interrogation subsiste sur la robustesse de la méthode, notamment en présence de petits échantillons.

6.2 Les indicateurs de Monte-Carlo

Concernant les indicateurs de Monte-Carlo, les résultats semblent plus souvent en faveur de la segmentation, en particulier pour les variables quantitatives et les taux de réponse élevés. En revanche, la méthode des scores s'en sort mieux en termes de biais quand le taux de répondants est plus faible. Ainsi, tous scénarii confondus, par rapport aux 294 cas étudiés, synthèse des 21 scénarii appliqués aux 14 variables d'intérêt, la segmentation permet d'améliorer le biais dans 52 % des cas par rapport à la méthode des scores, et surtout de diminuer l'erreur relative dans 59 % des cas.

Si globalement l'avantage se porterait plutôt sur la segmentation, dont le principal intérêt est de diminuer plus souvent l'erreur relative par rapport à la méthode des scores, il faut toutefois se garder de généraliser ce résultat, et ce, pour deux raisons. La première est que, même si les indicateurs de Monte-Carlo permettent de comparer les deux méthodes, les indicateurs restent tout de même assez proches. La seconde raison porte sur la variabilité des résultats en fonction des variables et du scénario.

Proportion de cas où la segmentation est meilleure selon l'indicateur de Monte-Carlo (en %, sur les 294 cas étudiés, soit 14 variables x 21 scenarii) :

		Biais relatif				Erreur relative			
		70 %	80 %	90 %	Total	70 %	80 %	90 %	Total
Variables quantitatives		**43**	**51**	**63**	**52**	**71**	**51**	**83**	**69**
dont	Aléatoire simple	20	40	20	27	100	80	100	93
	ZAU	100	100	100	100	100	100	100	100
	Taux d'endettement	60	60	60	60	100	100	60	87
	GRH	40	60	80	60	40	0	60	33
	GRH x taux d'endettement	40	60	80	60	40	20	60	40
	Secteur x taille	20	0	80	33	60	40	100	67
	Secteur x taille x taux d'endettement	20	40	20	27	60	20	100	60
Variables qualitatives		**44**	**48**	**62**	**51**	**62**	**51**	**63**	**59**
dont	Aléatoire simple	11	44	44	33	89	100	11	67
	ZAU	67	56	67	63	33	33	44	37
	Taux d'endettement	56	67	56	59	11	33	89	44
	GRH	67	44	56	56	67	67	67	67
	GRH x taux d'endettement	44	44	67	52	56	56	56	56
	Secteur x taille	33	44	78	52	44	44	78	56
	Secteur x taille x taux d'endettement	33	33	67	44	44	44	67	52
Ensemble des variables		**44**	**49**	**62**	**52**	**57**	**53**	**67**	**59**
dont	Aléatoire simple	14	43	36	31	93	93	43	76
	ZAU	79	71	79	76	57	57	64	60
	Taux d'endettement	57	64	57	60	43	57	79	60
	GRH	57	50	64	57	57	43	64	55
	GRH x taux d'endettement	43	50	71	55	50	43	57	50
	Secteur x taille	29	29	79	45	50	43	86	60
	Secteur x taille x taux d'endettement	29	36	50	38	50	36	79	55

Note : Les cases sont grisées lorsque la segmentation est meilleure dans la majorité des cas.

6.3 Synthèse et pistes d'approfondissement

Les indicateurs de Monte-Carlo étant finalement relativement proches, l'intérêt essentiel de la segmentation porte sur la phase de « modélisation ». La segmentation présente en effet plusieurs avantages. Elle facilite le regroupement des modalités, utilise davantage d'information auxiliaire et permet de caractériser les groupes de réponse homogènes constitués.

En marge de ces conclusions, il faut avoir à l'esprit que les deux méthodes de correction de la non-réponse ont été traitées dans un cadre particulier, celui de l'enquête TIC 2011, avec des choix parfois arbitraires sur leur mise en œuvre (nombre de groupes de réponse homogènes pour la

méthode des scores, critères d'arrêt pour la segmentation, etc.). Aussi, quelques pistes d'approfondissement peuvent être envisagées. D'autres tests pourraient être effectués pour analyser l'impact des différents paramètres (seuil de significativité, etc.). La stabilité des modèles de segmentation pourrait également être vérifiée par validation croisée ou par utilisation d'un échantillon d'apprentissage et d'un échantillon-test. Enfin, en pratique, la régression logistique pourrait être conservée en phase préliminaire, pour sélectionner les variables auxiliaires, avant d'utiliser la segmentation par arbres pour le regroupement des modalités et la formation des classes de repondération.

Bibliographie

Théorie des sondages :

- Ardilly, P. (2006). Les techniques de sondage, Technip.
- Le Guennec, J., and Sautory, O. (2005). Les sondages avec SAS, Insee/Cepe.

Correction de la non-réponse :

- Haziza, D. (2006). Traitement de la non-réponse dans les enquêtes, Ensai, support de cours FCDA.
- Caron, N. (2005). La correction de la non-réponse par repondération et par imputation, Insee, Document de travail, n° M0502.
- Neiter, B., and Buisson, B. (2010. Comment redresser une enquête thématique ?, Insee, Document de travail, n° E2010/01.

Régression logistique :

- Pommeret, D. (2008). Régression sur données catégorielles et sur données de comptage, Ensai, support de cours FCDA.
- Nakache, J.-P., and Confais, J. (2003). Statistique explicative appliquée, Technip.

Segmentation :

- Gelein, B. (2011). Méthodes de segmentation par arbres, Ensai, support de cours de 2e année.
- Costet, N. (2009). Méthode de segmentation par arbres binaires, Ensai, support de cours FCDA.
- Rakotomalala, R. (2005). Arbres de décision, Revue Modulad, n° 33.
- Nakache, J.-P., and Confais, J. (2003). Statistique explicative appliquée, Technip.
- Claudel, A., and Guevara, S., Utilisation des arbres de segmentation - Guide du chargé d'études pour CIS 2010, Insee Île-de-France, Document de travail.
- Tufféry, S. (2009). Étude de cas en statistique décisionnelle, Technip (macros SAS disponibles sous www.toeditions.com/Sources/tuffery_Etude-de-cas.htm).

Enquête TIC 2011 :

- L'enquête sur les technologies de l'information et de la communication, auprès des entreprises - TIC, Insee, Sources et méthodes (version du 17 janvier 2011).
- Demande d'expertise pour le tirage de l'échantillon pour l'enquête TIC 2011, Insee/DES, note interne n° /DG75-E430/ du 22 juillet 2010.
- Demande de tirage de l'échantillon pour l'enquête TIC 2011, Insee/DSE, note interne n° 601/DG75-E430/ du 2 novembre 2010.
- Tirage de l'échantillon pour l'enquête TIC 2011, Insee/UMSE, note interne n° 651/DG75-E101/AF du 22 novembre 2010.
- Redressement de l'enquête TIC 2011 (correction de la non-réponse et calage) - Note de cadrage, Insee/PISE, note interne du 21 avril 2011.

Annexes

Annexe 1 - Estimateurs

Cible et écart par rapport à la cible (en %) pour les variables quantitatives:

			Cible	Scores			Segmentation		
				70 %	80 %	90 %	70 %	80 %	90 %
A2	Nombre de personnes utilisant un ordinateur	SAS	5 890	0,0	0,0	0,0	0,0	0,0	0,0
		zau	5 890	-0,5	-0,4	-0,3	-0,5	-0,4	-0,3
		taux_endett	5 890	0,3	0,3	0,2	0,2	0,2	0,1
		GRH	5 890	0,1	-0,4	-0,8	0,2	0,1	0,1
		GRH_endett	5 890	0,3	-0,1	-0,6	0,3	0,2	0,2
		sect_taille	5 890	0,1	-0,1	-0,1	0,3	0,3	-0,1
		sect_taille_endett	5 890	0,0	-0,2	0,0	0,4	0,4	0,1
B4	Nombre de personnes utilisant Internet	SAS	4 680	0,0	0,0	0,0	0,0	-0,1	0,0
		zau	4 680	-0,6	-0,5	-0,4	-0,6	-0,5	-0,4
		taux_endett	4 680	0,2	0,3	0,2	0,2	0,2	0,1
		GRH	4 680	0,1	-0,4	-0,7	0,2	0,1	0,1
		GRH_endett	4 680	0,2	-0,2	-0,7	0,2	0,1	0,1
		sect_taille	4 680	0,1	-0,2	-0,1	0,1	0,3	-0,1
		sect_taille_endett	4 680	-0,1	-0,3	0,0	0,3	0,4	0,1
G2	Montant du CA généré via le web	SAS	70 093	0,2	0,3	0,1	0,2	0,3	0,1
		zau	70 093	-0,4	-0,2	-0,1	-0,4	-0,2	-0,1
		taux_endett	70 093	0,4	0,5	0,4	0,3	0,4	0,3
		GRH	70 093	0,3	-0,3	-0,7	0,3	0,2	0,0
		GRH_endett	70 093	1,1	0,7	0,1	1,0	1,1	0,8
		sect_taille	70 093	0,4	0,3	0,3	0,3	0,6	0,3
		sect_taille_endett	70 093	0,1	-0,2	-0,1	-0,2	0,0	-0,1
G5	Montant du CA généré via EDI	SAS	304 706	-0,1	0,0	0,0	-0,2	-0,1	0,0
		zau	304 706	-0,4	-0,2	-0,2	-0,3	-0,1	-0,2
		taux_endett	304 706	-0,1	-0,2	-0,2	-0,3	-0,4	-0,3
		GRH	304 706	0,2	-0,5	-0,9	0,8	0,6	0,3
		GRH_endett	304 706	-0,1	-0,7	-1,3	0,3	0,1	-0,2
		sect_taille	304 706	0,0	-0,1	-0,1	0,7	0,2	-0,1
		sect_taille_endett	304 706	-0,6	-0,8	-0,5	0,2	0,0	-0,6
G8	Montant des achats électroniques	SAS	266 397	-0,1	0,1	-0,1	-0,1	0,0	-0,1
		zau	266 397	-0,4	-0,4	-0,4	-0,4	-0,3	-0,4
		taux_endett	266 397	-0,4	-0,4	-0,3	-0,5	-0,5	-0,4
		GRH	266 397	0,8	0,4	0,0	0,5	0,5	0,1
		GRH_endett	266 397	0,6	0,1	-0,2	0,3	0,1	-0,1
		sect_taille	266 397	0,4	0,0	0,0	0,5	0,2	-0,1
		sect_taille_endett	266 397	0,0	-0,1	-0,1	-0,1	0,4	-0,3

Note : Montants en millions d'euros. En grisé, les cas où l'écart dépasse 0,5 %.

Cible (en %) et écart par rapport à la cible (en point de %) pour les variables qualitatives :

			Cible	Scores			Segmentation		
				70 %	80 %	90 %	70 %	80 %	90 %
B1	Présence d'un accès Internet	SAS	98,3	0,00	0,00	0,00	0,00	0,00	0,00
		zau	98,3	0,00	0,00	-0,01	0,00	0,00	0,00
		taux_endett	98,3	0,01	0,02	0,02	0,01	0,02	0,02
		GRH	98,3	0,02	0,02	0,02	0,03	0,02	0,02
		GRH_endett	98,3	0,00	0,01	0,01	0,02	0,02	0,01
		sect_taille	98,3	0,01	0,01	0,01	0,02	0,01	0,00
		sect_taille_endett	98,3	-0,04	-0,03	-0,02	-0,01	-0,02	-0,03
B6	Présence d'un site web ou d'une page d'accueil	SAS	62,7	-0,01	0,00	0,00	-0,02	-0,01	0,00
		zau	62,7	-0,09	-0,09	-0,06	-0,09	-0,09	-0,06
		taux_endett	62,7	0,08	0,09	0,09	0,07	0,08	0,08
		GRH	62,7	0,07	0,07	0,04	0,03	0,06	0,07
		GRH_endett	62,7	0,05	0,04	0,01	0,02	0,04	0,04
		sect_taille	62,7	0,01	-0,02	0,01	0,05	-0,05	-0,02
		sect_taille_endett	62,7	0,02	0,01	0,01	0,02	0,04	0,00
C1	Présence d'un système d'échange électronique traité automatiquement	SAS	48,1	0,00	0,01	-0,01	0,00	0,01	-0,01
		zau	48,1	-0,07	-0,06	-0,05	-0,07	-0,06	-0,04
		taux_endett	48,1	0,08	0,09	0,08	0,08	0,08	0,07
		GRH	48,1	0,08	0,08	0,06	0,07	0,07	0,04
		GRH_endett	48,1	0,17	0,13	0,11	0,18	0,13	0,09
		sect_taille	48,1	0,13	0,10	0,14	0,14	0,08	0,04
		sect_taille_endett	48,1	0,29	0,26	0,26	0,31	0,27	0,19
D1a	Factures électroniques aux clients par traitement automatique	SAS	9,6	-0,02	-0,01	0,00	-0,02	-0,01	0,00
		zau	9,6	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
		taux_endett	9,6	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
		GRH	9,6	0,03	0,02	0,00	0,02	0,02	0,00
		GRH_endett	9,6	0,02	0,00	-0,02	0,00	0,00	-0,02
		sect_taille	9,6	0,03	0,04	0,03	0,05	0,05	0,02
		sect_taille_endett	9,6	0,03	0,01	0,01	0,04	0,01	0,01
D1b	Factures électroniques aux clients par courrier ou pièces jointes pdf	SAS	34,8	-0,01	0,00	0,01	-0,01	0,00	0,01
		zau	34,8	0,01	0,03	0,03	0,01	0,03	0,02
		taux_endett	34,8	0,02	-0,01	-0,01	0,02	-0,01	0,00
		GRH	34,8	0,12	0,15	0,14	0,12	0,12	0,09
		GRH_endett	34,8	0,03	0,06	0,06	0,03	0,04	0,02
		sect_taille	34,8	0,09	0,09	0,08	0,07	0,08	0,02
		sect_taille_endett	34,8	0,11	0,11	0,09	0,12	0,13	0,05
G1	Réception de commandes de biens ou services sur le site web	SAS	9,3	0,00	0,00	0,00	0,00	0,00	0,00
		zau	9,3	-0,01	-0,02	-0,01	-0,01	-0,03	-0,01
		taux_endett	9,3	-0,03	-0,01	0,00	-0,03	-0,01	-0,01
		GRH	9,3	-0,03	-0,03	-0,05	-0,07	-0,04	-0,03
		GRH_endett	9,3	-0,04	-0,04	-0,05	-0,09	-0,06	-0,04
		sect_taille	9,3	-0,03	-0,04	-0,04	-0,07	-0,02	-0,01
		sect_taille_endett	9,3	-0,10	-0,11	-0,14	-0,19	-0,14	-0,11
G4	Réception de commandes de biens ou services via EDI	SAS	5,3	0,00	0,00	0,01	0,00	0,00	0,01
		zau	5,3	0,02	0,02	0,01	0,02	0,02	0,01
		taux_endett	5,3	0,02	0,02	0,02	0,02	0,01	0,01
		GRH	5,3	0,03	0,01	-0,01	0,04	0,04	0,02
		GRH_endett	5,3	0,03	0,00	-0,01	0,04	0,03	0,02
		sect_taille	5,3	0,02	0,02	0,02	0,07	0,02	0,02
		sect_taille_endett	5,3	0,01	-0,01	0,01	0,05	0,00	0,02
G7	Achat de biens et services par voie électronique	SAS	27,4	-0,01	0,01	0,00	-0,01	0,01	0,00
		zau	27,4	-0,06	-0,06	-0,03	-0,06	-0,06	-0,03
		taux_endett	27,4	0,09	0,08	0,08	0,08	0,07	0,07
		GRH	27,4	0,02	0,07	0,07	0,01	0,04	0,03
		GRH_endett	27,4	0,09	0,10	0,13	0,08	0,08	0,10
		sect_taille	27,4	0,12	0,06	0,08	0,10	0,03	0,01
		sect_taille_endett	27,4	0,29	0,26	0,24	0,25	0,25	0,19
H1	Utilisation d'outils fondés sur la RFID	SAS	2,5	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
		zau	2,5	-0,01	-0,01	-0,01	0,00	-0,01	-0,01
		taux_endett	2,5	-0,02	-0,01	-0,02	-0,02	-0,02	-0,02
		GRH	2,5	0,00	0,01	0,00	0,00	0,01	0,01
		GRH_endett	2,5	0,02	0,02	0,01	0,02	0,02	0,01
		sect_taille	2,5	-0,01	-0,01	-0,02	-0,01	-0,02	-0,02
		sect_taille_endett	2,5	0,02	0,01	0,02	0,02	0,02	0,01

Note : Proportion de ‘oui’, en %. En grisé, les cas où l’écart dépasse 0,05 point.

Annexe 2 - Indicateurs de Monte-Carlo

Rappel pour le biais relatif : $RB_{MC}\left(\hat{Y}^{meth}\right) = \frac{1}{R}\sum_{j=1}^{R}\frac{\hat{Y}_j^{meth} - Y}{Y} \times 100$ (en %)

Rappel pour le rapport des erreurs quadratiques moyennes : $\frac{MSE_{MC}\left(\hat{Y}^{seg}\right)}{MSE_{MC}\left(\hat{Y}^{score}\right)} = \frac{\frac{1}{R}\sum_{j=1}^{R}\left(\hat{Y}_j^{seg} - Y\right)^2}{\frac{1}{R}\sum_{j=1}^{R}\left(\hat{Y}_j^{score} - Y\right)^2}$

Pour faciliter la lecture, la qualité des résultats est indiquée à l'aide d'un code :

- biais relatif : les chiffres sont indiqués en gras et en grisé selon l'importance du biais,
- rapport des erreurs quadratiques moyennes : les cases sont en gras lorsque l'erreur est plus faible par la segmentation (< 1,0) et grisées lorsqu'elle est plus faible par la méthode des scores (> 1,0). Si la case est vide, le rapport vaut 1 (même performance pour les deux méthodes).

Indicateurs de Monte-Carlo pour les variables quantitatives :

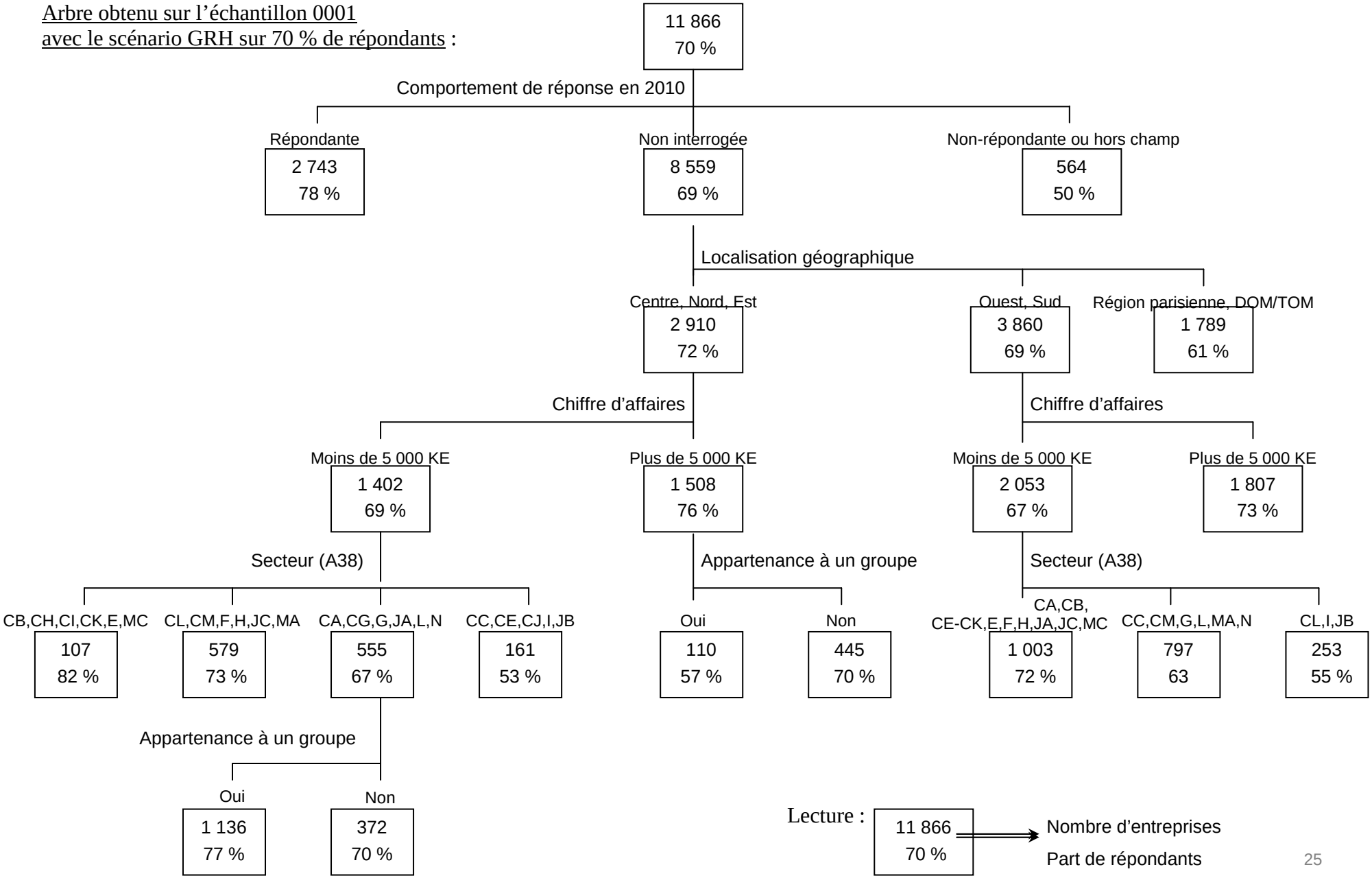
			Biais relatif RB_{MC} (en %)						**Rapport $MSE_{MC}^{seg}/MSE_{MC}^{score}$**		
			Scores			**Segmentation**					
			70 %	**80 %**	**90 %**	**70 %**	**80 %**	**90 %**	**70 %**	**80 %**	**90 %**
A2	Nombre de personnes utilisant un ordinateur	SAS	0,0	0,0	0,0	0,0	0,0	0,0	**0,91**	**0,95**	**0,94**
		zau	**-0,5**	-0,4	-0,3	**-0,5**	-0,4	-0,3	**0,85**	**0,87**	**0,93**
		taux_endett	0,3	0,3	0,2	0,2	0,2	0,1	**0,87**	**0,91**	**0,93**
		GRH	0,1	-0,4	**-0,8**	0,2	0,1	0,1	1,11	1,07	**0,66**
		GRH_endett	0,3	-0,1	**-0,6**	0,3	0,2	0,2	1,10	1,29	**0,85**
		sect_taille	0,1	-0,1	-0,1	0,3	0,3	-0,1	**0,99**	1,25	**0,95**
		sect_taille_endett	0,0	-0,2	0,0	0,4	0,4	0,1	1,12	1,27	**0,91**
B4	Nombre de personnes utilisant Internet	SAS	0,0	0,0	0,0	0,0	-0,1	0,0	**0,92**	**0,95**	**0,96**
		zau	**-0,6**	**-0,5**	-0,4	**-0,6**	**-0,5**	-0,4	**0,88**	**0,89**	**0,94**
		taux_endett	0,2	0,3	0,2	0,2	0,2	0,1	**0,90**	**0,93**	**0,96**
		GRH	0,1	-0,4	**-0,7**	0,2	0,1	0,1	1,10	1,06	**0,73**
		GRH_endett	0,2	-0,2	**-0,7**	0,2	0,1	0,1	1,10	1,17	**0,80**
		sect_taille	0,1	-0,2	-0,1	0,1	0,3	-0,1	**0,98**	1,16	**0,95**
		sect_taille_endett	-0,1	-0,3	0,0	0,3	0,4	0,1	1,04	1,17	**0,92**
G2	Montant du CA généré via le web	SAS	0,2	0,3	0,1	0,2	0,3	0,1	**0,99**		
		zau	-0,4	-0,2	-0,1	-0,4	-0,2	-0,1	**0,99**		
		taux_endett	0,4	**0,5**	0,4	0,3	0,4	0,3	**0,99**		**0,99**
		GRH	0,3	-0,3	**-0,7**	0,3	0,2	0,0		1,02	**0,98**
		GRH_endett	**1,1**	**0,7**	0,1	**1,0**	**1,1**	**0,8**	**0,98**	1,03	1,07
		sect_taille	0,4	0,3	0,3	0,3	**0,6**	0,3	**0,98**	1,01	
		sect_taille_endett	0,1	-0,2	-0,1	-0,2	0,0	-0,1	**0,99**		**0,98**
G5	Montant du CA généré via EDI	SAS	-0,1	0,0	0,0	-0,2	-0,1	0,0			**0,99**
		zau	-0,4	-0,2	-0,2	-0,3	-0,1	-0,2	**0,98**		**0,99**
		taux_endett	-0,1	-0,2	-0,2	-0,3	-0,4	-0,3	**0,99**	**0,99**	1,01
		GRH	0,2	**-0,5**	**-0,9**	**0,8**	**0,6**	0,3	1,06	1,08	1,01
		GRH_endett	-0,1	**-0,7**	**-1,3**	0,3	0,1	-0,2	1,01	1,01	**0,93**
		sect_taille	0,0	-0,1	-0,1	**0,7**	0,2	-0,1	1,03	**0,99**	
		sect_taille_endett	**-0,6**	**-0,8**	**-0,5**	0,2	0,0	**-0,6**	**0,99**	**0,97**	**0,99**
G8	Montant des achats électroniques	SAS	-0,1	0,1	-0,1	-0,1	0,0	-0,1	**0,99**		**0,99**
		zau	-0,4	-0,4	-0,4	-0,4	-0,3	-0,4	**0,99**		
		taux_endett	-0,4	-0,4	-0,3	**-0,5**	**-0,5**	-0,4		**0,99**	1,01
		GRH	**0,8**	0,4	0,0	**0,5**	**0,5**	0,1	**0,96**	1,01	1,02
		GRH_endett	**0,6**	0,1	-0,2	0,3	0,1	-0,1	**0,97**		1,02
		sect_taille	0,4	0,0	0,0	**0,5**	0,2	-0,1	1,01	**0,99**	**0,99**
		sect_taille_endett	0,0	-0,1	-0,1	-0,1	0,4	-0,3	**0,95**	1,02	**0,99**

Indicateurs de Monte-Carlo pour les variables qualitatives :

			Biais relatif RB_{MC}(en %)						Rapport $MSE_{MC}^{seg}/MSE_{MC}^{score}$		
			Scores			Segmentation					
			70 %	80 %	90 %	70 %	80 %	90 %	70 %	80 %	90 %
B1	Présence d'un accès Internet	SAS	0,0	0,0	0,0	0,0	0,0	0,0			
		zau	0,0	0,0	0,0	0,0	0,0	0,0			
		taux_endett	0,0	0,0	0,0	0,0	0,0	0,0			
		GRH	0,0	0,0	0,0	0,0	0,0	0,0	**0,99**	1,01	1,02
		GRH_endett	0,0	0,0	0,0	0,0	0,0	0,0		**0,99**	1,01
		sect_taille	0,0	0,0	0,0	0,0	0,0	0,0		1,01	
		sect_taille_endett	0,0	0,0	0,0	0,0	0,0	0,0	**0,96**	**0,99**	1,01
B6	Présence d'un site web ou d'une page d'accueil	SAS	0,0	0,0	0,0	0,0	0,0	0,0			
		zau	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1			1,01
		taux_endett	0,1	0,1	0,1	0,1	0,1	0,1			
		GRH	0,1	0,1	0,1	0,1	0,1	0,1			
		GRH_endett	0,1	0,1	0,0	0,0	0,1	0,1	**0,99**	**0,99**	
		sect_taille	0,0	0,0	0,0	0,1	-0,1	0,0	1,01		
		sect_taille_endett	0,0	0,0	0,0	0,0	0,1	0,0		1,01	
C1	Présence d'un système d'échange électronique traité automatiquement	SAS	0,0	0,0	0,0	0,0	0,0	0,0			
		zau	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	1,01		
		taux_endett	**0,2**	**0,2**	**0,2**	**0,2**	**0,2**	0,1			
		GRH	**0,2**	**0,2**	0,1	**0,2**	0,1	0,1		**0,99**	**0,98**
		GRH_endett	**0,4**	**0,3**	**0,2**	**0,4**	**0,3**	**0,2**	**0,99**	**0,99**	
		sect_taille	**0,3**	**0,2**	**0,3**	**0,3**	**0,2**	0,1	**0,99**	**0,99**	**0,96**
		sect_taille_endett	**0,6**	**0,6**	**0,5**	**0,7**	**0,6**	**0,4**	1,03	1,02	**0,95**
D1a	Factures électroniques aux clients par traitement automatique	SAS	**-0,2**	-0,1	0,0	**-0,2**	-0,1	0,0			
		zau	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1			
		taux_endett	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1			
		GRH	**0,3**	**0,2**	0,0	**0,2**	**0,2**	0,0	**0,99**	**0,99**	
		GRH_endett	**0,2**	0,0	**-0,2**	0,0	0,0	**-0,2**	**0,99**		
		sect_taille	**0,4**	**0,4**	**0,3**	**0,6**	**0,5**	**0,2**	1,01	1,01	
		sect_taille_endett	**0,3**	0,1	0,1	**0,5**	0,1	0,1		**0,99**	
D1b	Factures électroniques aux clients par courrier ou pièces jointes pdf	SAS	0,0	0,0	0,0	0,0	0,0	0,0			
		zau	0,0	0,1	0,1	0,0	0,1	0,1			**0,99**
		taux_endett	0,0	0,0	0,0	0,1	0,0	0,0			
		GRH	**0,3**	**0,4**	**0,4**	**0,3**	**0,3**	**0,3**		**0,98**	**0,98**
		GRH_endett	0,1	**0,2**	**0,2**	0,1	0,1	0,1	1,01	**0,99**	**0,99**
		sect_taille	**0,3**	**0,2**	**0,2**	**0,2**	**0,2**	0,1	**0,99**		**0,98**
		sect_taille_endett	**0,3**	**0,3**	**0,3**	**0,3**	**0,4**	0,1	1,01	1,02	**0,99**
G1	Réception de commandes de biens ou services sur le site web	SAS	0,0	0,1	0,0	0,0	0,1	0,0			
		zau	-0,1	**-0,3**	-0,1	-0,1	**-0,3**	-0,1			1,01
		taux_endett	**-0,3**	-0,1	0,0	**-0,3**	-0,1	-0,1			
		GRH	**-0,3**	**-0,4**	**-0,5**	**-0,8**	**-0,4**	**-0,3**	1,02	1,01	**0,98**
		GRH_endett	**-0,4**	**-0,4**	**-0,6**	**-1,0**	**-0,6**	**-0,4**	1,04	1,01	
		sect_taille	**-0,3**	**-0,4**	**-0,4**	**-0,7**	**-0,2**	-0,1	1,04		
		sect_taille_endett	**-1,1**	**-1,1**	**-1,5**	**-2,1**	**-1,5**	**-1,2**	1,17	1,03	**0,98**
G4	Réception de commandes de biens ou services via EDI	SAS	0,0	0,1	0,1	-0,1	0,1	0,1	**0,99**		
		zau	**0,3**	**0,3**	**0,2**	**0,4**	**0,3**	**0,2**		1,01	
		taux_endett	**0,4**	**0,3**	**0,3**	**0,3**	**0,2**	**0,2**		**0,99**	
		GRH	**0,6**	**0,2**	-0,1	**0,8**	**0,8**	**0,4**		1,04	1,02
		GRH_endett	**0,5**	0,1	**-0,2**	**0,7**	**0,6**	**0,3**		1,03	1,02
		sect_taille	**0,4**	**0,3**	**0,4**	**1,2**	**0,4**	**0,3**	1,07	**0,99**	
		sect_taille_endett	**0,2**	**-0,2**	**0,3**	**0,9**	0,0	**0,4**	1,02	**0,99**	1,02
G7	Achat de biens et services par voie électronique	SAS	0,0	0,0	0,0	0,0	0,0	0,0			
		zau	**-0,2**	**-0,2**	-0,1	**-0,2**	**-0,2**	-0,1			
		taux_endett	**0,3**	**0,3**	**0,3**	**0,3**	**0,3**	**0,3**			
		GRH	0,1	**0,3**	**0,3**	0,0	**0,2**	0,1	**0,99**	**0,99**	**0,99**
		GRH_endett	**0,3**	**0,4**	**0,5**	**0,3**	**0,3**	**0,4**		**0,99**	**0,98**
		sect_taille	**0,4**	**0,2**	**0,3**	**0,4**	0,1	0,1	**0,99**		**0,98**
		sect_taille_endett	**1,1**	**0,9**	**0,9**	**0,9**	**0,9**	**0,7**	**0,96**	**0,98**	**0,95**
H1	Utilisation d'outils fondés sur la RFID	SAS	**-0,5**	**-0,5**	**-0,4**	**-0,5**	**-0,5**	**-0,4**			
		zau	**-0,2**	**-0,3**	**-0,3**	**-0,2**	**-0,3**	**-0,3**			
		taux_endett	**-0,7**	**-0,6**	**-0,9**	**-0,8**	**-0,6**	**-0,9**		1,01	
		GRH	**0,2**	**0,4**	0,0	0,1	**0,5**	**0,2**	**0,97**	**0,99**	**0,99**
		GRH_endett	**1,0**	**0,7**	**0,3**	**0,9**	**0,8**	**0,5**	1,01	1,01	1,01
		sect_taille	**-0,3**	**-0,3**	**-0,7**	**-0,2**	**-0,7**	**-0,8**	**0,99**	**0,99**	
		sect_taille_endett	**0,7**	**0,5**	**0,6**	**0,7**	**0,8**	**0,3**	**0,98**	1,01	**0,98**

Annexe 3 - Exemple d'arbre de segmentation

Arbre obtenu sur l'échantillon 0001
avec le scénario GRH sur 70 % de répondants :



Annexe 4 - L'enquête TIC 2011

L'enquête annuelle sur l'utilisation des technologies de l'information et de la communication (TIC) et le commerce électronique est réalisée par l'Insee en partenariat avec le service de l'observation et des statistiques (SOeS) du ministère en charge du développement durable et le service de la statistique et de la prospective (SSP) du ministère chargé de l'agriculture. Elle s'inscrit dans le dispositif d'enquêtes européennes, en application du règlement européen n°1006/2009 du 16 septembre 2009 amendant le règlement du 21 avril 2004.

Les objectifs de l'enquête

L'enquête vise à mieux connaître la diffusion des TIC dans les entreprises. Elle cherche notamment à apprécier la place des outils nouveaux dans les relations externes de l'entreprise (internet, commerce électronique) et dans leur fonctionnement interne (réseaux, systèmes intégrés de gestion).

Elle est composée d'un tronc commun de questions articulées autour de trois thèmes principaux : l'équipement en TIC, l'accès et l'usage d'internet, le commerce électronique, auxquels s'ajoute chaque année un module, déterminé par le règlement européen annuel d'application du règlement cadre, traitant d'un thème nouveau ou approfondissant l'un des thèmes de base. Pour l'enquête 2011, le module supplémentaire porte sur l'utilisation des technologies fondées sur l'identification par radio-fréquence (RFID) (cf. questionnaire).

Le champ de l'enquête

À partir de l'enquête TIC 2009, le champ couvre les entreprises marchandes exploitantes de 10 salariés et plus de la métropole, appartenant aux secteurs suivants de la Naf rév. 2 :

- l'industrie manufacturière (section C)
- la production et la distribution d'électricité, de gaz, de vapeur et d'air conditionné (section D)
- la production et la distribution d'eau, l'assainissement, la gestion des déchets et la dépollution (section E)
- la construction (section F)
- le commerce, la réparation d'automobiles et de motocycles (section G)
- les transports et l'entreposage (section H)
- l'hébergement et la restauration (section I)
- l'information et la communication (section J)
- les activités immobilières (section L)
- les activités spécialisées, scientifiques et techniques hors activités vétérinaires (divisions 69 à 74)
- les activités de services administratifs et de soutien (section N)
- la réparation d'ordinateurs et d'équipements de communication (groupe 95.1).

Le plan de sondage

La base de sondage est le répertoire Sirene (Système informatique pour le répertoire des entreprises et de leurs établissements).

L'échantillon est stratifié par secteur d'activité et par taille, les « strates de sondage » étant définies par le croisement des modalités. Les modalités des tranches d'effectifs sont au nombre de cinq (10 à 19 salariés, 20 à 49 salariés, 50 à 249 salariés, 250 à 499 salariés, 500 salariés et plus). Les modalités des secteurs d'activité ont des niveaux d'agrégation très divers (de la classe au regroupement de sections), mais il peut y avoir des regroupements de divisions ou de groupes.

Le nombre d'entreprises à échantillonner diffère selon les strates. Les entreprises de plus de 500 salariés sont interrogées exhaustivement (sauf pour l'activité de nettoyage, qui comporte beaucoup de grandes entreprises). Pour les strates de taille d'effectif inférieure, le nombre d'entreprises à interroger a été obtenu par le biais d'une allocation proportionnelle au nombre de salariés.

Au final, selon l'année d'enquête, environ 13 000 entreprises sont interrogées.

Sources : Présentation de l'enquête sur Insee.fr (L'enquête sur les technologies de l'information et de la communication, auprès des entreprises - TIC, Insee, Sources et méthodes, janvier 2011) ; notes n° 651/DG75-E101/AF, 601/DG75-E430/ et /DG75-E430 sur le tirage de l'échantillon de l'enquête TIC 2011.

TIC 2011

Enquête sur les Technologies de l'Information et de la Communication et le Commerce électronique

Enquête conduite, dans le cadre de la Statistique publique, par l'Institut national de la statistique et des études économiques **(Insee)**, le Service de l'observation et des statistiques **(SOeS)**, le Service de la statistique et de la prospective **(Ssp)**.

*Vu l'avis favorable du Conseil national de l'information statistique, cette **enquête**, reconnue d'intérêt général et de qualité statistique, est **obligatoire**, visa n°2011 ?????? du ministre de l'Économie, de l'Industrie et de l'Emploi valable pour 2011. Aux termes de l'article 6 de la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques, les renseignements transmis en réponse au présent questionnaire ne sauraient en aucun cas être utilisés à des fins de contrôle fiscal ou de répression économique. La loi n° 78-17 du 6 janvier 1978 relative à l'in formatique, aux fichiers et aux libertés, s'applique aux réponses faites à la présente enquête par les entreprises individuelles. Elle leur garantit un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès de l'Insee.*

Pour plus de renseignements, vous pouvez contacter à l'Insee :
«GestNom»
Tél : «GestTel» - Fax : «GestFax»

Merci d'adresser votre réponse avant le : «D_Ech»
à : Insee Midi-Pyrénées
36, rue des Trente-Six-Ponts
BP 94217
31054 TOULOUSE CEDEX 4

Nom de l'entreprise : «NL1_NOMEN» - «NL2_COMP» SIREN : «SIREN»
Code APE : «APEL» Libellé de l'activité : «LIB_APEL»
Adresse : «NL3_CADR» - «NL4_VOIE» - «NL5_DISP» - «NL6_CODEPOST» - «NL7_LOCALITE»

CACHET DE L'ENTREPRISE
«NL1_NOMEN» - «NL2_COMP»
«NL3_CADR»
«NL4_VOIE»
«NL5_DISP»
«NL6_CODEPOST»
«NL7_LOCALITE»

Nom et coordonnées de la personne répondant à ce questionnaire :
Mme/Mlle/M : «C_NOM»
Fonction : «C_FONC»
Téléphone : «C_TEL»
Fax : «C_FAX»
Courriel : «C_COURRIEL1» @ «C_COURRIEL2»
Adresse (si différente de celle de l'entreprise) :
«C_L1_NOMEN» - «C_L2_COMP» - «C_L3_CADR» - «C_L4_VOIE»
«C_L5_DISP» - «C_L6_CODEPOST» - «C_L7_LOCALITE»

Site Web de **l'entreprise interrogée** : http://www. «ADR_SITEWEB»

Ce questionnaire concerne **votre entreprise en tant qu'entité juridique (tous établissements confondus), à l'exclusion de toute autre entité** (groupe ou établissement secondaire). Il peut concerner **plusieurs responsables** de la direction générale ou du service informatique. Cette enquête a pour objectif d'évaluer l'importance de la diffusion et de l'utilisation des technologies de l'information et de la communication (TIC) dans les entreprises. Menée chaque année dans le cadre d'une investigation européenne, elle permet de recueillir des points de repère et de comparaison importants pour l'orientation de l'action publique.

Module A : Utilisation d'ordinateurs et de réseaux

A1 **Votre entreprise utilise-t-elle au moins un ordinateur ?** ..... **(question filtre)** — OUI / NON
Le terme « ordinateur » inclut les PC, les nettops*, les ordinateurs portables (*notebooks**...), les assistants numériques personnels (*PDA**) ou les téléphones intelligents* (*smartphone*…) ..... A1_ORDINATEUR
Si NON passer à X1

A2 Parmi les personnes employées dans votre entreprise, dont le nombre doit être indiqué à la question **X1a**, combien utilisent un ordinateur au moins une fois par semaine ? ..... A2_EMP_ORDI_VAL
Si vous ne pouvez pas fournir ce nombre, veuillez indiquer leur proportion parmi les personnes employées. ..... A2_EMP_ORDI_PCT

A3 Votre entreprise utilise-t-elle des logiciels « libres* » (*encore dits logiciels « open source »*) dans les cas suivants ? — OUI / NON
a) un système d'exploitation (*Linux*…) ..... A3a_LIBRE_SYSTEXP
b) un logiciel de bureau (*OpenOffice*…) ..... A3b_LIBRE_LOGIC
c) un ERP ou PGI* Open source pour automatiser les procédures d'entreprises (*OpenERP, Compiere, ERP5*…) ..... A3c_LIBRE_ERP

Module B : Accès et utilisation d'internet (*champ : entreprises avec ordinateurs*)

B1 **Votre entreprise a-t-elle un accès à internet* ?** ..... **(question filtre)** — OUI / NON — B1_ACCES_INTERNET
Si NON passer à C1

B2 Votre entreprise a-t-elle les types de connexion à internet suivants ?
a) Modem traditionnel (*accès commuté sur une ligne téléphonique normale*) ou connexion RNIS* (*en anglais ISDN*) ..... B2a_TEL_ANALOGIQUE
b) Connexion DSL* (*xDSL, ADSL, SDSL…*) ..... B2b_DSL
c) Autre connexion* fixe à internet (*câble, accès sans fil au réseau fixe, ligne louée… [relais de trame, Ethernet métropolitain, CPL…]*) ..... B2c_AUTR_CONNEX_FIXES
d) Connexion haut débit mobile* avec au moins une technologie 3G (*UMTS, CDMA2000, 1Xevdo, HSDPA…*) ..... B2d_CONNEX_MOB_3G
e) Autre connexion mobile* (*téléphone mobile analogique, GSM, GPRS, EDGE…*) ..... B2e_AUTR_CONNEX_MOB

B3 Quelle est la vitesse de téléchargement maximale contractuelle de la connexion internet la plus rapide de votre entr
cochez une seule case

a) Moins de 2	b) De 2 à moins de 10	c) De 10 à moins de 30	d) De 30 à moins de 100	B3_VITESSE_CONNEX
☐	☐	☐	☐	☐

B4 Parmi les personnes employées dans votre entreprise, dont le nombre doit être indiqué à la question **X1a**, combien utilisent au moins une fois par semaine un ordinateur **avec un accès à internet** ? ..... B4_EMP_INTERNET_VAL
Si vous ne pouvez pas fournir ce nombre, veuillez indiquer leur proportion parmi les personnes employées . ..... B4_EMP_INTERNET_PCT

B5 Parmi les personnes employées dans votre entreprise, dont le nombre doit être indiqué à la question **X1a**, combien disposent d'un **appareil portable** dédié à l'entreprise, **ayant au moins une technologie 3G*** pour accéder à internet ? ..... B5_EMP_PORTABLE3G_VAL
Si vous ne pouvez pas fournir ce nombre, veuillez indiquer leur proportion parmi les personnes employées. ..... B5_EMP_PORTABLE3G_PCT

B6 **Votre entreprise a-t-elle un site Web* ou une page d'accueil* ?** ..... **(question filtre)** — OUI / NON — B6_SITE_WEB
Si NON passer à B8

B7 Le site ou la page d'accueil de votre entreprise propose-t-il actuellement les services suivants ?
a) la commande ou la réservation en ligne (*« shopping cart*», caddie virtuel…*) ..... B7a_COMMANDE_LIGNE
b) des catalogues et/ou des listes de prix de biens ou services ..... B7b_CATALOGUES

UTILISATION D'INTERNET EN RELATION AVEC LES AUTORITÉS PUBLIQUES* (*champ : entreprises avec accès à internet*) — OUI / NON

B8 En 2010, votre entreprise a-t-elle utilisé internet ?
a) pour obtenir de l'information sur les sites web ou les pages d'accueil des autorités publiques ..... B8a_INFORMATION
b) pour obtenir des formulaires sur les sites web ou les pages d'accueil des autorités publiques (*déclaration d'impôts* ..... B8b_FORMULAIRES
c) pour retourner électroniquement un formulaire rempli (*formulaires de déclaration de douane ou de TVA…*) ..... B8c_RENS_FORMUL
d) pour le traitement « tout électronique » d'une procédure administrative (*déclaration, enregistrement, demande d'autorisation…*) ..... B8d_FORMAL_ADMIN

B9 En 2010, votre entreprise a-t-elle utilisé internet pour gérer les procédures administratives suivantes ?
(*en retournant un formulaire rempli électroniquement*)
a) déclaration de cotisations sociales pour les employés ..... B9a_COTIS_SOC
b) déclaration d'impôts sur les sociétés ..... B9b_IMPOTS
c) déclaration de TVA ..... B9c_TVA
d) déclaration de droits de douane / contributions indirectes ..... B9d_DOUANE

B10 Pensez-vous que certaines des raisons suivantes limitent les relations électroniques de votre entreprise avec les autorités publiques ?
a) inquiétude vis-à-vis de la confidentialité et de la sécurité ..... B10a_SECURITE
b) certaines procédures électroniques sont trop compliquées et/ou demandent trop de temps ..... B10b_TEMPS
c) certaines procédures électroniques requièrent toujours un échange de courrier ou des visites de personnes ..... B10c_COURRIER
d) ne connaît pas les possibilités de procédures électroniques ..... B10d_PROC_INCONNUES

B11 En 2010, votre entreprise a-t-elle utilisé internet pour accéder à une information sur des documents et des spécifications d'appel d'offre dans le cadre de marché électronique des autorités publiques*? ..... B11_INFO_APPEL_OFFRE

B12 En 2010, votre entreprise a-t-elle utilisé internet pour proposer des biens ou des services dans le cadre de marché électronique des autorités publiques (*dépôt d'appel d'offre électronique**) ?
a) en France ..... B12a_DEPOT_OFFRE_FR
b) dans un autre pays de l'Union Européenne ..... B12b_DEPOT_OFFRE_UE
Si OUI à B12a ou à B12b, passer à C1

B13 En 2010, votre entreprise n'a pas proposé de biens ou services dans le cadre de marché électronique des autorités publiques (*dépôt d'appel d'offre électronique*) : est-ce parce qu'elle ne vend pas au secteur public ? ..... B13_VENTES_ADMIN

Module C : Envoi et réception de messages appropriés à des traitements automatiques vers ou en provenance de systèmes extérieurs (*champ : entreprises avec ordinateurs*)

« Transmission électronique et traitement automatique de l'information » signifie :
*– **envoi et/ou réception de messages** (commande, facture, opération de paiement, description de produits, document de transport, déclaration d'impôts…) ;*

C2 Si oui, pour quelles opérations parmi les suivantes ? — OUI / NON

a) envoyer des instructions de paiement à des institutions financières ..... C2a_PAIEMENT
b) envoyer ou réceptionner des informations sur les produits (*catalogues, listes de prix…*) ..... C2b_INFO_PROD
c) envoyer ou réceptionner des documents de transport (*bordereaux d'expédition…*) ..... C2c_TRANSPORT
d) envoyer ou réceptionner des données avec des autorités publiques (*déclarations de chiffre d'affaires, données statistiq… déclarations d'importation ou d'exportation...*) C2d_DONNEES_ADMIN

Module D : Facturation électronique (*champ : entreprises avec ordinateurs*)

Facturation électronique signifie envoi ou réception de facture :
- *dans un format **qui permet son traitement automatique** (EDI*,XML ou formulaire web…) ;*
  *ou dans un format **requérant une action humaine** (courriel avec pièce jointe en pdf…) ;*
- *vers ou en provenance d'autres entreprises, d'autorités publiques, d'institutions financières ou de particuliers ;*
- *par le site web du vendeur, la banque du client ou d'autres voies électroniques.*

D1 Votre entreprise **envoie**-t-elle à des clients des factures électroniques*? — OUI / NON

a) dans une structure standardisée appropiée pour leur traitement automatique (*EDI,XML…*) ..... D1a_CLIENTS_FACT_AUTO
b) **ne permettant pas** de traitement informatique (*courriel, pièce jointe au format pdf…*) D1b_CLIENTS_FACT_NONAUT

D2 Votre entreprise **reçoit**-elle des fournisseurs des factures électroniques dans une structure standardisée appropriée pour leur traitement automatique (*EDI, UBL, XML…*) ? ..... D2_FOURNISSEURS_FACT_AUTO

Module E : Partage automatique de l'information au sein de l'entreprise (*champ : entreprises avec ordinateurs*)

Partage de l'information électroniquement ou automatiquement entre différents services de votre entreprise signifie au moins une des propositions suivantes :
- *utilisation d'une seule application pour assurer les différentes fonctions de l'entreprise (PGI ou ERP*…) ;*
- *liaison (chaînage) des données entre les applications qui assurent les différentes fonctions de l'entreprise ;*
- *utilisation d'une base de données commune ou d'un entrepôt de données partagé accessible par les applications qui assurent les différentes fonctions de l'entreprise ;*
- *à l'intérieur de l'entreprise, envoi et réception par voie électronique d'informations pouvant être traitées de manière automatique.*

E1 Quand votre entreprise reçoit des bons de commande (*par voie électronique ou non*), l'information est-elle partagée électroniquement ou automatiquement à l'aide d'un logiciel dédié aux services suivants ? — OUI / NON

a) votre gestion des niveaux de stocks ..... E1a_COM_STOCK
b) votre comptabilité ..... E1b_COM_COMPTA
c) votre gestion de production (*ou de services*) ..... E1c_COM_PROD
d) votre gestion de la distribution ..... E1d_COM_DISTRIB

E2 Lorsque votre entreprise effectue des ordres d'achat (*par voie électronique ou non*), l'information est-elle partagée électroniquement ou automatiquement à l'aide d'un logiciel dédié aux fonctions suivantes ?

a) votre gestion des niveaux de stocks ..... E2a_ACHAT_STOCK
b) votre comptabilité ..... E2b_ACHAT_COMPTA

E3 Votre entreprise a-t-elle utilisé un progiciel de gestion intégré (*PGI ou ERP**) pour partager l'information entre les différents pôles de l'entreprise (*comptabilité, finance, planning, production, marketing…*) ? ..... E3_PGI_ERP — Ne sait pas

E4 Votre entreprise utilise-t-elle une application pour la gestion de la relation client* (*aussi appelée CRM*) qui lui permet de :

a) collecter, conserver et rendre accessible à d'autres services l'information clientèle ? ..... E4a_INFO_CLIENT
b) analyser l'information clientèle à des fins de marketing (*fixation des prix, organisation des ventes promotionnelles, choix des canaux de distribution…*) ? ..... E4b_MARKETING

E5 Votre entreprise utilise-t-elle les outils informatiques suivants ?

a) outils de travail collaboratifs (*groupware*, vidéoconférence**…*) ..... E5a_VIDEOCONF
b) outils de modélisation et d'automatisation (*Workflow*, BPMS**…*) ..... E5b_MODELISATION
c) outils de CAO interne (*Conception Assistée par Ordinateur*) ..... E5c_CAO_INTERNE
d) outils de CAO collaborative (*entre entreprises*) ..... E5d_CAO_COLLABO

Module F : Les TIC et l'impact environnemental (*champ : entreprises avec ordinateurs*)

F1 Votre entreprise a-t-elle mis en place l'une des procédures suivantes : — OUI / NON

a) procédures visant à la réduction des impressions papier et des photocopies ? ..... F1a_REDUC_IMP
b) procédures visant à la réduction de la consommation énergétique dans l'utilisation de votre équipement TIC ? (*consigne au personnel pour éteindre les ordinateurs et les écrans, utilisation de matériel d'arrêt automatique des équipements T… utilisation de matériel périphérique multifonctions [imprimante, scanner, photocopieur]*) ..... F1b_REDUC_ENERGIE_TIC
c) utilisation du téléphone, d'internet ou de visioconférence au lieu de déplacement ..... F1c_REDUC_DEPLT

F2 Votre entreprise a-t-elle mis en place une quelconque application de technologie de l'information **dédiée** à reduire la consommation d'énergie des processus d'activité ? (*y compris l'optimisation des procédures de travail, des processu… production, du transport ou de la logistique*) ..... F2_REDUC_ENERGIE_PROCESS

Module G : Le commerce électronique* (*champ : entreprises avec ordinateurs*)

⊙ VENTES PAR COMMERCE ÉLECTRONIQUE

VENTES PAR WEB

G1 **En 2010, votre entreprise a-t-elle reçu des commandes de biens ou services qui ont été passées sur le site Web de votre entreprise ?** (*à l'exclusion des courriels saisis manuellement*) ..... **(question filtre)** — OUI / NON
Les ventes sur le Web sont des ventes effectuées via un magasin en ligne (*webshop*) ou via les formulaires Web sur le site Internet de votre entreprise ou à l'extranet, indépendamment de la façon dont le web est accessible [*ordinateur, portable, téléphone mobile…*] G1_VENTES_WEB
Si NON passer à G4

G2 Quel a été en 2010 le montant du chiffre d'affaires hors taxes généré par des commandes **reçues** qui ont été passées sur le site Web de votre entreprise ? (*en milliers d'euros*) ..... G2_VENT_WEB_VAL

*Si vous ne pouvez pas indiquer le montant, veuillez en donner une estimation en % du chiffre d'affaires total hors taxes de votre entreprise (indiqué à la question **X1b**)* ..... G2_VENT_ WEB _PCT

G3 En 2010, votre entreprise a-t-elle **reçu** des commandes qui ont été passées sur le site Web de votre entreprise par des clients localisés dans les zones géographiques suivantes ? — OUI / NON

a) en France ..... G3a _VENTE_WEB_FR
b) dans les autres pays de l'Union Européenne ..... G3b _VENTE_WEB_UE
c) dans le reste du monde ..... G3c_VENTE_WEB_RDM

VENTES DE TYPE EDI*

G4 **En 2010, votre entreprise a-t-elle reçu des commandes de biens ou services qui ont été passées via des messages de type EDI ?** ..... **(question filtre)** — OUI / NON G4_VENTES_EDI
Si NON passer à G7

G5 Quel a été en 2010 le montant du chiffre d'affaires hors taxes généré par des commandes **reçues** qui ont été passées via des messages de type EDI ? (*en milliers d'euros*) ..... G5_VENT_EDI_VAL

*Si vous ne pouvez pas indiquer le montant, veuillez en donner une estimation en % du chiffre d'affaires total hors taxes de votre entreprise (indiqués à la question **X1b**)* ..... G5_VENT_EDI_PCT

G6 En 2010, votre entreprise a-t-elle **reçu** des commandes qui ont été passées via des messages de type EDI par des clients localisés dans les zones géographiques suivantes ?

a) en France ..... G6a_VENTE_EDI_FR
b) dans les autres pays de l'Union Européenne ..... G6b_VENTE_EDI_UE
c) dans le reste du monde ..... G6c_VENTE_EDI_RDM

⊙ ACHATS PAR COMMERCE ÉLECTRONIQUE

G7 **En 2010, votre entreprise a-t-elle passé des commandes de biens ou services via les réseaux informatiques ?** (*via un site web ou un message de type EDI mais en excluant les courriels saisis manuellement*) ..... **(question filtre)** — OUI / NON G7_ACHATS
Si NON passer à H1

G8 Quel a été en 2010 le montant des achats hors taxes générés par des commandes **passées** via les réseaux informatiques ? (*en milliers d'euros*) ..... K€ G8_ACHAT_VAL

*Si vous ne pouvez pas indiquer le montant, veuillez en donner une estimation en % du total des achats hors taxes de l'entreprise (indiqués à la question **X1c**)* ..... G8_ACHAT_PCT

G9 En 2010, votre entreprise a-t-elle **passé** des commandes via un site web ou via des messages de type EDI auprès de fournisseurs localisés dans les zones géographiques suivantes ?

a) en France ..... G9a_COM_FR
b) dans les autres pays de l'Union Européenne ..... G9b_ COM_UE
c) dans le reste du monde ..... G9c_ COM_RDM

Module H : Utilisation des technologies basées sur l'Identification par Radio Fréquence (RFID*) (*champ : entreprises avec ordinateurs*)

Identification par Radio Fréquence (en anglais, Radio Frequency Identification RFID) désigne une méthode d'identification automatique pour stocker et récupérer des données à distance en utilisant les étiquettes RFID ou transpondeurs.
Une étiquette RFID est un dispositif qui peut être attaché ou incorporé à l'objet et qui transmet des données via les ondes radio.

H1 Votre entreprise utilise-t-elle des outils basés sur la technologie RFID ? ..... **(question filtre)** — OUI / NON H1_UTILISATION_RFID
Si NON passer à X1

H2 Dans quels buts votre entreprise utilise-t-elle la technologie RFID ?

a) l'identification des personnes ou le contrôle des accès ..... H2a_IDENT_PERS
b) le suivi et le contrôle de la production industrielle, la chaîne d'approvisionnement et du suivi des stocks, le service, l'entretien ou la gestion d'actifs (*dans le cadre du processus de production et de prestation de services*) ..... H2b_SUIVI_PROCESS
c) l'identification de produits après-vente (*contrôle des vols, contrefaçon, information sur les allergènes*) ..... H2c_IDENT_PROD

Module X : Données de cadrage

X1 Données caractéristiques de la dimension de l'entreprise pour 2010

a) effectif annuel moyen* en 2010 (*en nombre de personnes occupées*) ..... X1a_EFF _ nombre
X1b_CA
X1c_ACHAT