

## MODÈLE POUR PRÉDIRE LA PROBABILITÉ D'AVOIR UNE LIMITATION D'ACTIVITÉ RELIÉE À L'EMPLOI À PARTIR DE L'ENQUÊTE SUR LA SANTÉ ET LES LIMITATIONS D'ACTIVITÉ

Daniel Hurtubise et Eric Langlet<sup>1</sup>

### RÉSUMÉ

L'Enquête sur la Santé et les Limitations d'Activités (ESLA) a eu lieu en 1986 et 1991, mais n'a pas pu être menée en 1996 par manque de budget. Développement des ressources humaines Canada voudrait quand même produire des tableaux portant sur les limitations d'activités pour leur programme d'Équité en matière d'emploi. Les estimations des probabilités d'être limité seront basées sur un modèle reliant l'indicateur d'équité en matière d'emploi de limitation d'activité contenu dans l'ESLA 1991 aux variables du recensement de 1991. Ce modèle sera alors appliqué sur les données du recensement de 1996 afin d'obtenir une probabilité d'être limité pour chaque individu de l'échantillon 2B. Différentes méthodes pour prédire les probabilités d'être limité sont proposées et comparées, soit l'analyse par arbre de classification, l'analyse par régression logistique, l'analyse discriminante, et une méthode de proportion.

MOTS-CLÉS : Limitation d'activité; analyse par arbre de classification; régression logistique; comparaison de modèles.

### ABSTRACT

Health and Activity Limitation Survey (HALS) was held in 1986 and 1991, but not in 1996 due to budget constraints. Human Resources Development Canada would like to produce tables of persons limited in their activity for their Employment Equity Program. Estimates of the probability of being limited will be based on a model connecting the 1991 HALS employment equity disability indicator with the 1991 Census variables. This model will then be applied to the 1996 Census data to obtain a probability of being disabled for each person in the 2B sample. Different methods to predict probabilities of being disabled are proposed and compared, namely the classification tree analysis, the logistic regression analysis, the discriminant analysis, and a proportion method.

KEY WORDS : Activity Limitation; Classification Tree Analysis; Logistic Regression; Model Comparison.

### 1. INTRODUCTION

L'objectif de ce projet est d'estimer les comptes d'individus étant limités dans leurs activités (selon la définition donnée plus bas), en vue de produire les tableaux requis par Développement des ressources humaines Canada (DRHC) dans le cadre du programme d'équité en matière d'emploi (PEME).

Antérieurement, ces tableaux étaient produits suite à l'enquête post-censitaire sur la Santé et les Limitations d'Activité (ESLA), qui a eu lieu en 1986 et 1991. Une enquête post-censitaire identifie la population cible de l'enquête selon les réponses fournies à certaines questions du questionnaire long

du recensement (aussi appelé formulaire 2B). Cette enquête n'a pas eu lieu en 1996; cependant, DRHC voudrait quand même produire des tableaux portant sur les limitations d'activités.

La limitation d'activité, définie par le PEME, se décrit comme suit (définition de 1991): la personne interrogée est limitée si elle a entre 15 et 64 ans, si elle a indiqué qu'elle avait une limitation dans ses activités selon la définition de limitation de l'ESLA et si elle est limitée dans le travail qu'elle fait ou qu'elle pourrait faire ou si elle pense que son employeur ou son employeur potentiel pourrait la considérer limitée dans ses activités.

Le sigle *EEPWD* (*Employment Equity Person With*

<sup>1</sup> Daniel Hurtubise et Eric Langlet, Division des méthodes d'enquêtes sociales, 15e étage, édifice RH Coats, Statistique Canada, Ottawa, Ontario, K1A 0T6, hurtdan@statcan.ca, langlet@statcan.ca

*Disability*) sera utilisé pour décrire les personnes qui sont limitées selon cette définition.

La façon d'atteindre cet objectif consiste à étudier la relation entre la limitation d'activité (provenant de *EEPWD*) et les variables du recensement: il s'agit en fait de déterminer quelles variables du recensement expliquent le plus la limitation d'activité et quelle est la relation entre ces variables et la limitation d'activité selon *EEPWD*. Il ne restera qu'à appliquer ce modèle aux individus faisant partie de l'échantillon 2B de 1996. Le résultat de cette étude sera la détermination d'une probabilité d'être limité, qui combinée avec le poids final de chaque individu du 2B, donnera une estimation du nombre de personnes limitées. On ne cherche pas à déterminer qui sera limité et qui ne le sera pas.

Dans le présent article, l'ESLA sera abordée dans la section 2 alors que la section 3 présentera les données ainsi que les poids d'échantillonnage utilisés. Les différentes méthodes d'analyse et les mesures de comparaison seront présentées dans la section 4. Les résultats obtenus seront décrits dans la section 5. L'estimation de la variance sera brièvement abordée dans la section 6, et la section 7 conclura cet article.

## 2. L'ENQUÊTE SUR LA SANTÉ ET LES LIMITATIONS D'ACTIVITÉ

L'ESLA est une enquête post-censitaire qui a été menée en 1986 et 1991. Le plan d'échantillonnage de cette enquête est conçu à partir des réponses aux questions du recensement (formulaire 2B) portant sur les limitations d'activité. À l'aide de ces questions, on identifie quatre variables de limitation au recensement: limitation à la maison, à l'école ou au travail, à long terme et autres types de limitations. Si une personne s'est déclarée limitée à au moins une des quatre catégories, alors cette personne est limitée selon le recensement. Dans le but d'identifier à l'avance une portion importante de la population des personnes avec une incapacité selon l'ESLA, un échantillon avec une grande fraction de sondage parmi les gens s'étant déclarés limités au recensement a été tiré. Cependant, les enquêtes précédentes ont montré que le groupe des personnes limitées selon l'ESLA n'est pas complètement couvert par les personnes limitées au recensement. En conséquence, pour éliminer tout biais possible résultant d'une sous-couverture de la population, une fraction dix fois moindre a été utilisée pour sélectionner les gens n'ayant pas déclaré de limitation au recensement. La

limitation au recensement a donc servi de variable de stratification à l'enquête.

Cette enquête, qui permet d'obtenir de l'information sur la nature et la sévérité des incapacités, est divisée en deux composantes: composante des ménages et celle des institutions. Deux questionnaires sont utilisés selon l'âge des répondants: un questionnaire adulte pour les 15 ans et plus et un questionnaire pour les enfants. Les questionnaires retenus pour cette étude sont ceux de la composante ménage pour lesquels les répondants ont entre 15 et 64 ans. À partir de cette enquête, on peut définir deux indicateurs d'incapacité: celui de l'ESLA et celui du PEME, soit *EEPWD* tel que défini dans la section précédente. Parmi les personnes limitées à l'ESLA, on en retrouve près de la moitié qui se sont déclarées non limitées au recensement; une proportion semblable d'individus étant limités selon *EEPWD* se sont déclarés non limités au recensement. Ces données montrent bien que les définitions de limitation d'activité sont différentes selon le recensement et selon PEME.

## 3. DESCRIPTION DES DONNÉES

Les données utilisées proviennent du croisement entre le fichier du recensement de 1991 et du fichier de l'ESLA de 1991, et ce pour les personnes âgées entre 15 et 64 ans, hors institution. Le fichier résultant contient 86 220 observations et 66 variables. Les observations pour lesquelles certaines variables avaient une valeur manquante ont été imputées, la plupart de temps en utilisant le mode de la distribution de chaque variable. La variable dépendante est l'indicatrice de *EEPWD*, qui a la valeur 1 si la personne est limitée selon *EEPWD* et 0 sinon. Les variables indépendantes sont les variables du recensement.

### 3.1 Analyse préliminaire des variables

Afin de réduire le nombre de variables, des tests du Chi-deux et de Student ont été utilisés pour identifier les variables non significatives en fonction de la limitation d'activité selon *EEPWD*. Ces tests sont sensibles à la taille d'échantillon; certaines variables ayant été déclarées significatives ont été retirées de l'étude de par leur faible valeur de la statistique.

Dans le but d'optimiser le modèle, une analyse graphique a été réalisée. Chaque variable a été séparée selon les déciles de sa distribution, et la proportion de personnes limitées a été calculée par décile. Pour que la modélisation logistique soit efficace, on devrait retrouver une relation linéaire entre la moyenne de

chaque décile et la valeur logit de la proportion de personnes limitées par décile. Si non, une transformation de la variable s'impose. Cette transformation peut être soit une transformation continue ou une catégorisation de variables continues. Seule la variable *age* a été conservée comme variable continue, et nous avons ajouté le carré de la variable *age* dans le modèle pour la strate limitée au recensement. Les autres variables continues ont été catégorisées, soit de façon dichotomique ou polytomique. Suite à ces analyses, 33 variables ont été conservées dans le fichier.

### 3.2 Analyse des individus

Une analyse sommaire des individus, selon qu'ils soient limités ou non au recensement, montrent une grande différence de profil entre les deux groupes. En général, la sévérité de la limitation selon *EEPWD* des personnes n'étant pas limitées au recensement est beaucoup moindre que celle des personnes limitées au recensement. Nous allons donc les étudier séparément, en développant un modèle pour chacun de ces groupes. Il est à noter que pour cette étude les individus n'ayant pas répondu aux questions sur les limitations d'activité au recensement sont mis dans le même groupe que les individus non limités au recensement, ce qui correspond à la stratification du plan de sondage.

### 3.3 Poids d'échantillonnage

Chaque individu de l'ESLA étant sélectionné par échantillonnage, un poids est attaché à chacun de ceux-ci reflétant le nombre de personnes que l'individu représente dans la population. Des procédures statistiques pondérées seront donc considérées pour analyser ces données. De façon à se servir de la taille de l'échantillon dans le calcul des degrés de liberté plutôt que de la taille de la population, les poids ont été standardisés de sorte que la somme de ces derniers soit égale à la taille de l'échantillon. Également, des poids tenant compte de l'effet de plan de l'enquête sont aussi considérés. L'effet de plan de l'enquête ESLA est estimé à 2. Pour le calcul de ces derniers poids, les poids standardisés sont divisés par l'effet de plan. Cette approche étant plus conservatrice que la méthode pondérée précédente, moins de variables seront déclarées significatives dans le modèle.

Les différentes méthodes présentées dans la section suivante sont développées en utilisant trois sortes de poids, soient les poids standardisés, les poids standardisés avec effet de plan ainsi que des poids de 1 (approche non pondérée).

## 4. MÉTHODES D'ANALYSE ET MESURES COMPARATIVES

Plusieurs méthodes d'analyse sont possibles pour expliquer la limitation d'activité en fonction des variables du recensement. Dans notre cas, nous nous attarderons aux trois méthodes suivantes: analyse par arbre de classification, régression logistique, méthode de proportion. Le résultat de chacune de ces méthodes sera l'obtention d'un modèle qui nous permettra de calculer la probabilité d'être limité pour tous les individus. L'analyse discriminante a été considérée mais les résultats préliminaires obtenus ont démontré que ce type d'analyse n'est pas tellement recommandé avec des variables de type catégorique, celles-ci ayant été dichotomisées avant d'utiliser l'analyse discriminante.

### 4.1 Analyse par arbre de classification

Cette méthode crée un arbre, selon les variables du recensement, qui permet de classer les individus dans des sous-groupes les plus homogènes possibles relativement à la limitation d'activité selon *EEPWD*, et les plus hétérogènes possibles entre les sous-groupes.

Le logiciel utilisé, KnowledgeSeeker, offre deux méthodes différentes pour analyser les données: la méthode exhaustive et la méthode "cluster". Voici un bref résumé de ces deux méthodes.

#### 4.1.1 Méthode exhaustive

Le test utilisé est un test du Chi-deux basé sur le croisement de la limitation d'activité selon *EEPWD* et chacune des variables du recensement. Les variables continues sont préalablement groupées en classes. Le logiciel tente de regrouper les niveaux des variables de façon à maximiser la statistique du  $\chi^2$ . Toutes les combinaisons possibles de regroupement de niveaux sont testées par variable, seul le regroupement qui donne la statistique la plus élevée est conservé. Certains critères d'arrêt en terme de taille (pour la formation d'un groupe et la division d'un groupe) et en terme de niveau de signification ont été utilisés. On assigne aux individus de chaque feuille terminale la même probabilité d'être limité, qui correspond à la proportion d'individus étant limités dans cette feuille.

#### 4.1.2 Méthode "cluster"

Cette méthode cherche essentiellement à maximiser les différences entre les sous-groupes par rapport à la variable de limitation d'activité selon *EEPWD*. Elle est plus rapide que la méthode précédente, et a

tendance à produire des sous-groupes plus naturels. Cette méthode, de type "CHAID" (*Chi-square Automatic Interaction Detection*), maximise le niveau de signification d'une statistique  $\chi^2$  à chaque partition. Contrairement à la méthode exhaustive, cette méthode ne teste pas tous les regroupements possibles de niveaux. Une fois que des niveaux sont regroupés et qu'ils ne sont pas séparés dans l'étape subséquente, ils ne peuvent être considérés avec d'autres niveaux, même si ces autres regroupements donneraient une statistique du  $\chi^2$  supérieure.

La méthode "cluster" est celle qui a été retenue pour notre analyse, pour une raison d'économie (rapidité d'exécution) ainsi que le fait que les sous-groupes créés sont plus naturels que ceux de la méthode exhaustive.

Pour appliquer le modèle obtenu aux individus du recensement de 1996, on répartit ces personnes dans chacune des feuilles de l'arbre, selon leurs caractéristiques au recensement, et on leur assigne la probabilité d'être limité (calculée avec les données de 1991) correspondant à la feuille terminale. Donc tous les individus qui appartiennent à la même feuille auront la même probabilité d'être limité.

#### 4.2 Modèle logistique

Une sélection préliminaire de variables a été faite en utilisant la procédure *probit* du logiciel SAS. Une caractéristique de cette procédure est qu'elle teste les différents niveaux des variables catégoriques par rapport à un des niveaux, qui est appelé le niveau de référence. On détermine ainsi les niveaux qui sont significativement différents du niveau de référence, en fonction de la limitation d'activité selon *EEPWD*. Les niveaux qui ne sont pas significativement différents du niveau de référence ont été combinés avec ce dernier pour créer notre nouveau groupe de référence. Les niveaux significativement différents du niveau de référence ont été croisés avec la variable dépendante pour tester s'ils sont significativement différents entre eux. Certains de ces niveaux ont été regroupés. Ensuite, une variable dichotomique par niveau significatif ou groupe de niveaux significatifs a été créée; ces variables sont utilisées par la procédure *logistic*. L'interprétation des variables retenues dans le modèle se fera par rapport au groupe de référence. Ensuite, la méthode de sélection à rebours (« *backward* ») a été choisie, qui élimine les variables les moins significatives du modèle. Afin de s'assurer d'une meilleure interprétation des paramètres du modèle, toutes les variables dichotomiques relatives à une seule variable catégorique ont été forcées dans le modèle dès qu'une de ces variables dichotomiques

était significative. Une étude des interactions a aussi été réalisée. Comme il y a beaucoup de variables dans l'étude, nous nous sommes attardés à l'interaction entre l'âge et les quatre variables de limitation au recensement parce que ces variables expliquent le plus la limitation d'activité selon *EEPWD*.

Une analyse des diagnostics a été réalisée, selon les mesures établies par Pregibon (1981) et disponibles via le logiciel SAS. Les conclusions ne permettent pas d'éliminer des observations car aucune de celles-ci ne se démarquent réellement des autres. Parmi les individus ayant un grand résidu (valeur réelle moins valeur prédite), les individus limités selon *EEPWD* ont des caractéristiques au recensement semblables aux individus non limités selon *EEPWD*, et les individus non limités selon *EEPWD* ont des caractéristiques au recensement semblables aux individus limités selon *EEPWD*.

#### 4.3 Méthode des proportions

Cette méthode, contrairement aux autres décrites dans cet article, ne considère que les variables utilisées dans les tableaux finaux. Nous ne cherchons pas de relation entre la variable dépendante et les variables du recensement à l'aide d'un modèle statistique. Dans le but d'obtenir des estimations cohérentes entre les tableaux, nous avons créé un tableau global, à partir des variables de chaque tableau requis par DRHC. Comme ce tableau comporte beaucoup de cellules, on a procédé à un regroupement de niveaux de variables, selon la proportion d'individus limités pour chaque niveau d'une variable, et à un regroupement de cellules, celles dans lesquelles on retrouve très peu de personnes, afin de permettre une meilleure estimation des proportions. La proportion estimée est donc, pour chaque cellule obtenue, la proportion de personnes limitées à *EEPWD* par rapport au nombre total de personnes dans chaque cellule. Pour les prédictions, on assigne aux individus du fichier 2B du recensement de 1996 qui sont dans le regroupement de cellules la proportion correspondante à la cellule correspondant à leur caractéristique au recensement. L'hypothèse sous-jacente à cette méthode est que l'on considère que la proportion de personnes limitées à *EEPWD* est constante dans le temps, à l'intérieur de chaque cellule du tableau.

#### 4.4 Mesures de comparaison entre les méthodes

Dans le but de pouvoir comparer les trois méthodes, nous avons mis en place des mesures statistiques permettant de comparer les résultats obtenus de chaque méthode. Il y a deux types de mesure: mesure de *précision* et mesure de *cohérence*. Pour le calcul

des différentes mesures, les résultats obtenus avec les individus limités au recensement et ceux obtenus avec les individus non limités au recensement ont été combinés, afin d'obtenir une mesure globale de chaque méthode. Le principe sous-jacent aux mesures est le suivant: une bonne méthode pour l'estimation de la probabilité d'être limité dans ses activités selon *EEPWD* devrait donner des probabilités près de 0 (ou faibles) pour les gens non limités et près de 1 (ou fortes) pour les gens limités.

#### 4.4.1 Mesures de précision

La mesure de précision vient vérifier la valeur des probabilités: est-ce que les gens qui ne sont pas limités ont bien une probabilité près de 0? est-ce que les gens qui sont limités ont bien une probabilité près de 1?

Cette mesure calcule l'erreur de prédiction. Elle consiste à donner une *pénalité* à chaque individu, pondérée par son poids final. La pénalité consiste, pour les individus non limités selon *EEPWD*, à leur probabilité d'être limité; pour les individus limités selon *EEPWD*, elle consiste à leur probabilité de ne pas être limité. Ainsi, si la méthode est bonne, la valeur de la mesure sera près de 0, puisque la probabilité d'être limité sera faible pour les individus non limités et forte pour les individus limités. Le défaut de cette mesure est qu'elle est linéaire, ceci a pour conséquence que des distributions fort différentes de probabilité d'être limité donnent des valeurs de mesure très semblables. Dans le but d'améliorer cette mesure, la pénalité sera élevée au carré (similaire au score de Brier, mesure analogue à la somme des carrés résiduelle en régression appliquée à une variable binaire). Dans notre cas, nous avons utilisé des mesures pondérées et avons additionné les deux composantes de la mesure. Le fait de pondérer séparément chacune des deux parties de la mesure non pondérée et de la mesure pondérée selon la limitation à *EEPWD* est de s'assurer que l'on donne autant de poids aux individus des deux groupes, sinon il y aurait un avantage manifeste à bien classer les individus non limités à *EEPWD*, étant donné qu'ils forment environ 93% de la population. Le but est de minimiser la valeur de ces mesures.

Une autre mesure de l'erreur d'ajustement est celle du logarithme de vraisemblance de Bernoulli (*Bernoulli log-likelihood*). Cette mesure est plus appropriée que la mesure de Brier lorsque la variable dépendante est binaire et suit une loi de Bernoulli. Le logarithme de la probabilité prédite intervient dans cette formule, et sachant que le logarithme naturel d'une valeur comprise entre 0 et 1 est négatif, la valeur de cette

mesure sera donc négative. On cherche à maximiser la valeur de cette mesure, on choisira donc la méthode pour laquelle la mesure est la plus près de 0.

#### 4.4.2 Mesures de cohérence

Cette mesure calcule la force de dépendance entre les variables explicatives et la variable réponse (dichotomique dans notre cas). Si la méthode est bonne, les individus non limités devraient avoir leur probabilité d'être limité plus petite que les individus limités (d'où l'appellation de cohérence). La mesure que nous utilisons est celle du *D de Somer*. Chaque individu non limité est apparié avec un individu limité. On appelle cette paire concordante si la probabilité d'être limité est plus faible pour l'individu non limité; elle est discordante si l'inverse est observé; elle est égale si les deux probabilités sont égales (à 0.002 près).

La mesure utilise la différence entre le nombre de paires concordantes et le nombre de paires discordantes, divisée par le nombre total de paires. Une bonne méthode donnera plus de paires concordantes que discordantes, d'où une mesure près de 1, la valeur maximale du *D de Somer*. La valeur 1 sera réalisée s'il n'y a pas de chevauchement entre la distribution des probabilités prédites pour chaque groupe. Une statistique semblable au *D de Somer* est aussi utilisée, soit la statistique *c*, fournie par le logiciel SAS dans sa procédure *logistic*.

#### 4.4.3 Répétitions

Pour évaluer la capacité de prédiction d'un modèle statistique, il est bon de le valider sur un échantillon test indépendant de l'échantillon ayant servi à développer le modèle (échantillon d'apprentissage). Ceci est dû au fait que le modèle s'ajuste toujours mieux aux données ayant servi à le créer que sur des observations indépendantes, introduisant ainsi un biais. On peut répéter ce processus sur plusieurs échantillons d'apprentissage et plusieurs échantillons tests. Dans ce but, nous diviserons notre échantillon en quatre quarts de la façon suivante: chaque échantillon d'apprentissage contiendra le trois-quart des données, qui serviront à estimer le modèle, et le quart restant servira d'échantillon test (ou échantillon de validation). En alternant chaque quart, on obtient quatre modèles différents (cette méthode est connue sous le nom de "4-fold cross validation"). La validation consiste à comparer les différentes valeurs prises par les mesures décrites dans cette section.

Il est à noter que la division de l'échantillon en quatre quarts a été faite en utilisant un plan d'échantillonnage

aléatoire simple parmi les personnes limitées à *EEPWD* et celles non limitées à *EEPWD*. Ceci assurera une représentation égale des personnes limitées et des personnes non limitées dans chaque quart.

## 5. RÉSULTATS

En utilisant les mesures définies dans la section précédente, les résultats obtenus avec la régression logistique sont meilleurs que ceux obtenus avec l'arbre de classification et l'analyse discriminante; il est difficile de dire, en se basant seulement sur les mesures présentées dans la section précédente, si la régression logistique est meilleure que la méthode de proportion. Les différences observées sont relativement faibles dans certains cas. Le modèle non pondéré semble meilleur que les modèles pondérés, avec ou sans effet de plan. De plus, le modèle de régression logistique avec interaction de variables est meilleur que le modèle sans interaction.

Afin de départager les méthodes de régression logistique et celle des proportions, un test a été effectué sur un échantillon systématique de 1 individu sur 10 à partir du fichier du questionnaire 2B du recensement de 1991. Il s'agissait de recréer un des tableaux produits en 1991 à l'aide des deux modèles, et de comparer les résultats obtenus avec le tableau produit en 1991. Le modèle logistique est celui qui a produit les meilleurs résultats. En conséquence, la méthode qui sera retenue pour estimer les comptes d'individus étant limités dans leur activité selon *EEPWD* est la régression logistique pondérée avec termes d'interaction.

## 6. ESTIMATION DE LA VARIANCE

L'estimation de la variance se fera pour chaque cellule de chaque tableau; ainsi le coefficient de variation sera calculé pour chaque cellule. La variance sera calculée en tenant compte des deux plans de sondage impliqués: celui de l'ESLA de 1991 et celui du questionnaire long du recensement de 1996. Le plan de l'ESLA est un plan à deux degrés stratifié, alors que celui du questionnaire long du recensement de 1996 est un plan stratifié systématique avec une fraction de sondage de 1 ménage sur 5 pour tout le pays. Deux approches sont considérées: soit une approche par le *jackknife* ou une approche de type de linéarisation de Taylor.

## 7. CONCLUSION

L'hypothèse sous-jacente au modèle retenu est que la relation entre l'indicateur d'incapacité de *EEPWD* et les variables du recensement est constante dans le temps. Le modèle pondéré est conservé même s'il produit des résultats légèrement moins bons que le modèle non pondéré pour éviter un biais dans les estimations des paramètres, sachant que les données proviennent d'un sondage. Comme il semble qu'il y aura une ESLA en 2001, on peut voir les estimations calculées selon la méthode présentée dans cet article comme des estimations inter-censitaires, qui ne peuvent en aucun cas se comparer à celles qui auraient été obtenues par une véritable enquête en 1996.

## REMERCIEMENTS

Les auteurs aimeraient remercier, pour leur participation à ce projet : Georgia Roberts, Maryse Rivoal et Eric Lesage, et pour la révision de l'article : Anne-Marie Houle et Martin Lachance.

## RÉFÉRENCES

- BIGGS D., DE VILLE B., SUEN E. (1991). A method of choosing multiway partitions for classification and decision trees, *Journal of Applied Statistics*, vol 18, pp 49-62
- DENIS J., DUFOUR J., GRONDIN C., LAVIGNE M., LYNCH J., MORIN J.-P (1993). *Méthodologie de l'enquête sur la santé et les limitations d'activités 1991, composante ménages*
- KASS G.V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, vol 29, pp 119-127
- PREGIBON D. (1981). Logistic regression diagnostics, *The Annals of Statistics*, Vol. 9, no 4, pp. 705-724
- SAS Institute (1997). *Data mining using SAS Entreprise Miner*
- Statistique Canada (1993). *Enquête sur la santé et les limitations d'activités, 1991: guide de l'utilisateur*