Bulletin de Méthodologie Sociologique
2016, Vol. 132 26–43

DOI: 10.1177/0759106316662589
bms.sagepub.com

Modèles de mélange fini (FMM) appliqués à la segmentation du marché du travail à Bogota

Thibaud Deguilhem
GREThA UMR-CNRS 5113, Pessac, France, et CEDE, Bogota, Colombia

Suneha Seetahul
GREThA UMR-CNRS 5113, Pessac, France

Abstract

Finite Mixture Models (FMM) Applied to the Segmentation of the Bogota Labor Market. The purpose of this article is to demonstrate that the semi-parametric approach with Finite Mixture Models (FMM) is a real alternative to clustering procedures to explain social phenomena. We present the general framework of FMM before detailing the three stages of Finite Mixture Regression Models (FMRM). An empirical application of this new method is used to analyze the segmentation between formal and informal economy on the labor market in Bogota.

Résumé

Le propos de cet article est de démontrer que l'approche semi-paramétrique des modèles de mélange fini (FMM) constitue une véritable alternative aux procédures de *clustering* pour expliquer les phénomènes sociaux. Nous présentons le cadre général des FMM avant de détailler les trois étapes successives des modèles de mélange de régressions (FMRM). Une application empirique de cette méthode originale est développée afin d'analyser la segmentation entre économie formelle et informelle sur le marché du travail à Bogota.

Keywords

Finite Mixture Models (FMM), Semi-parametric Methods, Labour Market Segmentation

Corresponding Author:
Thibaud Deguilhem, GREThA UMR-CNRS 5113, Avenue Léon Duguit, Pessac, 33600, France
Email: thibaud.deguilhem@u-bordeaux.fr

Mots clés
Modèles de mélange fini, Méthode semi-paramétrique, Segmentation du marché du travail

Introduction

Dans le champ des sciences sociales, l'identification et la description de différents groupes homogènes au sein d'une population donnée est une étape primordiale, en particulier pour de nombreuses études traitant de la segmentation du marché du travail (Piore, 1983; Boston, 1990; Gindling, 1991; Leontaridi, 1998; Combarnous et Labazée, 2002; Hudson, 2007). Il devient alors nécessaire d'adopter une approche méthodologique qui permette aux résultats de refléter au mieux cette fragmentation dans la distribution des données. En ce sens, les modèles de mélange fini possèdent des attributs intéressants (Adams, 2016).

Les méthodes de classification hiérarchiques ou non-hiérarchiques représentent encore aujourd'hui les instruments les plus utilisés en sciences sociales pour l'indentification de groupes homogènes (Wedel et al., 2000, Wedel et Kamakura, 2000). Pourtant, face à l'éclatement des réalités et des pratiques sociales qui nécessite une démonstration rigoureuse d'une potentielle hétérogénéité, ces outils apparaissent insuffisants pour définir statistiquement le nombre "optimal" de groupes d'individus qui peuvent composer un même ensemble et estimer des effets différenciés (Salem et Bensidoun, 2012). En l'occurrence, le nombre de regroupements va être défini *ex ante* et sera intimement lié à la méthode de découpage retenue (en particulier pour les procédures supervisées). Dans ce sens, Lubke et Muthen (2005: 23) notent que la méthode de *clustering k-means* est fondée sur le choix discrétionnaire d'un critère qui consiste à minimiser la variance au sein de chaque groupe et à maximiser la variance entre les différents groupes.

Les Finite Mixture Models (FMM), ou modèles de mélange fini, sont depuis le début des années 2000 considérés comme la principale alternative aux algorithmes de classification (Stahl et Sallis, 2012; Tuma et Decker, 2013). D'une grande utilité dans de nombreux domaines où la diversité est la règle et l'unicité l'exception, les FMM se caractérisent par une certaine flexibilité statistique dans l'analyse de l'hétérogénéité entre groupes au sein d'une même population statistique. De plus, Andrews et al. (2010) décrivent cet outil comme la meilleure alternative puisqu'elle opère à l'aide d'un modèle statistique formel.

Bien que ces outils se soient propagés dans différents champs disciplinaires allant de la biologie moléculaire à l'astrophysique (Deb et Trivedi, 1997; Tuma et Decker, 2013), peu d'applications en sciences sociales usent de ces modèles afin d'estimer des paramètres inconnus sur des populations hétérogènes tout en déterminant la probabilité pour les individus d'appartenir à une sous-population (Keane et Wolpin, 1997; Conway et Deb, 2005; Deb et Trivedi, 1997, 2002, 2013; Deb et al., 2011). Venant appuyer la pertinence de l'illustration que nous avons choisie, quelques variantes de ces modèles FMM connaissent de nombreux développements et extensions en économie du travail (Cunha et al., 2010; Adams, 2016), en particulier pour éclairer l'hétérogénéité des formes

d'emploi dans les pays en développement (PED) (Salem et Bensidoun, 2012; Günther et Launov, 2012).

Ainsi, le propos de cet article est de démontrer comment nous pouvons revisiter le traitement de l'hétérogénéité entre différents groupes afin d'obtenir une meilleure connaissance des phénomènes étudiés. Dans un premier temps, nous proposons un retour sur les différentes techniques de *clustering* avant d'entamer une présentation générale des modèles de mélange fini. Dans un second temps, nous détaillons les trois étapes successives des modèles de mélange de régressions (FMRM) avant de présenter une application empirique sur des données en coupe transversale issues d'une enquête ménage colombienne[1]. Cette illustration nous permet de démontrer l'intérêt de cette méthode dans le traitement de la segmentation existante entre économie formelle et informelle[2] à Bogota. Dans un dernier temps, nous mettons en perspective les modèles de mélange fini en présentant la portée et les limites empiriques de ces méthodes.

Les modèles de mélange fini

Des méthodes de classification standards à la classification par modélisation statistique. Il existe quatre grands types d'outils statistiques permettant d'assigner des individus au sein de groupes homogènes. Le premier type regroupe les méthodes de partitionnement multivarié non-supervisées appelées classification hiérarchique (CH). Elles utilisent la notion de distance (*dissimilarité*) entre les observations prises deux à deux, et vont ainsi réaffecter ces dernières au sein de différents groupes homogènes les plus distincts les uns des autres (Lebart et al., 2006). Ces outils produisent alors un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Cet arbre représente une hiérarchie de partitionnement, et l'utilisateur doit alors choisir une partition adéquate en "tronquant" le dendrogramme de classification au niveau souhaité. Cependant, cette troncature fait intervenir une décision discrétionnaire quant au nombre de groupes en fonction de la question posée par l'utilisateur (Stahl et Sallis, 2012). De plus, ces méthodes ne sont pas facilement utilisables sur de grandes bases de données.

Le deuxième groupe de méthodes non-supervisées rassemble les outils connus sous le nom d'analyse exploratoire multivariée dont les plus populaires en sciences sociales sont l'Analyse en Composantes Principales (ACP) et l'Analyse en Composantes Multiples (ACM) (Lebart et al., 2006). Ces méthodes peuvent être considérées comme des outils de projection permettant de visualiser les observations depuis un espace à p dimensions des p variables vers un espace à k dimensions ($k < p$) tel qu'un maximum d'information soit conservée (l'information est ici mesurée au travers de la variance totale du nuage de points). Pourtant, ces méthodes rencontrent certaines difficultés quant à la détermination du nombre de dimensions à retenir. Bien que des techniques[3] permettent d'objectiver la sélection du nombre de facteurs, elles sont peu contraignantes et souvent mises en relation pour confirmer ce qui semblait liée à la question posée par l'utilisateur (Tuma et Decker, 2013). Enfin, les mêmes difficultés liées à la non-supervision affectent ces méthodes. Tous ces outils non-supervisés peuvent être utilisés de façon complémentaire afin de réduire les difficultés évoquées (Le Guen et

Jaffeux, 1989; Deguilhem et Frontenaud, 2016), mais sans jamais les surpasser totalement (Tuma et Decker, 2013).

Le troisième groupe de méthodes réunit les algorithmes supervisés et non-hiérarchiques, appelés nuées dynamiques[4], consistant à regrouper des individus à travers un nombre de groupes prédéterminés en optimisant une fonction fondée sur un critère de similarité entre les individus (Lebart et al., 2006). Cette méthode constitue une alternative aux outils non-supervisés évoqués plus haut et reste souvent utilisée en sciences sociales (Combarnous et Labazée, 2002). Toutefois, l'inconvénient de cette méthode est qu'elle ne permet pas de découvrir quel peut être un nombre "optimal" de classes, ni de visualiser la proximité entre les classes ou les observations (Everitt et al., 2011).

Le dernier type d'outils regroupe des méthodes de modélisation statistique dont font partie les modèles de mélange fini (Li et al., 2007). Ces derniers ont pour intérêt considérable de remédier aux deux grandes limites posées par les méthodes de classification. En effet, les typologies effectuées avec ces outils ne reposent pas sur des modèles statistiques formels mais plutôt sur une démarche itérative et intuitive qui consiste à regrouper les individus les plus semblables possibles et éloigner les individus les plus différents. Ainsi la détermination du nombre de groupes est soumise à une décision souvent discrétionnaire sur le choix critère de similarité et *dissimilarité* permettant ces regroupements et leur nombre. Pour ces raisons, il apparait intéressant d'utiliser les méthodes modélisation statistique telles que les FMM qui reposent sur des lois statistiques (telles que les distributions Gaussiennes ou de Poisson) et permettent d'objectiver la caractérisation des regroupements grâce à la comparaison de critères d'information statistique des différentes partitions possibles au sein de la population.

Caractérisation des modèles FMM. Ces modèles apparaissent particulièrement utiles pour modéliser l'hétérogénéité d'une population dans l'estimation d'une fonction de densité de probabilité (FDP) et au sein de modèles de régression (ces derniers sont aussi appelés *Finite Mixture Regression Models* (FMRM), voir p.7). Cela implique que la FDP de cet ensemble peut être appréhendée par le mélange des FDP associées à des groupes distincts d'individus. Dans le cadre des modèles de régression, l'objectif est alors d'expliquer la valeur d'une variable dépendante (ou endogène) par un vecteur de caractéristiques (variables exogènes) dont la combinaison de valeurs sera différente selon le groupe homogène identifié. Autre spécificité, dans ces modèles de mélange les sous-groupes auxquels appartiennent les individus ne sont pas connus et doivent être inférés, nous captons alors une éventuelle *hétérogénéité inobservée*.

"Pour l'estimation d'une densité de probabilité, un modèle de mélange définit la densité de la population comme une combinaison de densités paramétriques" (Ahamada et Flachaire, 2008: 99). Ainsi, nous pouvons écrire :

$$f(y,\theta) = \sum_{c=1}^{C} \pi_c f_c(y;\theta_c) \tag{1}$$

où θ est l'ensemble des paramètres, C le nombre de sous-groupes distincts, appelé aussi nombre de composantes du mélange,π_c la proportion de la population appartenant au

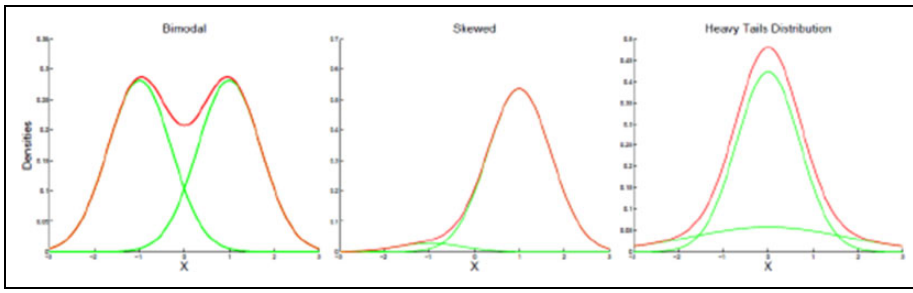

Figure 1. Exemples de mélange fini de lois normales
Note : Dans le graphique 1, les densités représentées ont le même poids $\pi_1 = \pi_2 = 0; 5$, et une même variance $\sigma_1 = \sigma_2 = 1$. Les graphiques 2 et 3 traduisent des proportions différentes et des variances inégales

sous-groupe c et $f_c(y, \theta_c)$ une distribution de probabilité caractérisée par un ensemble de paramètres contenus dans θ_c. Les paramètres π_c ont les propriétés suivantes :

$$0 \leq \pi_c \leq 1$$

$$\sum_{c=1}^{C} \pi_c = 1$$

Le rôle de l'estimation consiste à déterminer les paramètres inconnus, en considérant la vraisemblance de chacune des observations à appartenir à chacun des sous-groupes. Ainsi, les FMM permettent de spécifier différentes régressions sur les groupes distincts pour une même variable expliquée, et d'identifier des effets différenciés de certains facteurs explicatifs.

En définitive, les FMM constituent une méthode semi-paramétrique (estimateur non-paramétrique de la fonction de densité), entre l'estimation paramétrique et l'estimation non-paramétrique par noyau. L'aspect paramétrique se reflète par le fait qu'une fonction paramétrique est exprimée pour chaque composante, l'aspect non-paramétriques se caractérise par la présence d'un nombre inconnu de composantes (si C = 1, le mélange se réduit à une seule fonction paramétrique, si C = n, le mélange se ramène à une estimation d'une fonction de densité par la méthode du noyau).

Mélange fini de lois normales. Selon la théorie des modèles de mélange, n'importe quelle densité de probabilité peut être estimée de façon convergente par un mélange gaussien. Ce cas de figure est d'ailleurs largement observé dans les études empiriques (Ghosal et Van der Vaart, 2001; Davidson et Flachaire, 2007; Cowell et Flachaire, 2007). La Figure 1 présente plusieurs mélanges de deux distributions normales, dont la densité peut s'écrire de façon générique :

$$\pi_{c=1}\phi(y; \mu_{c=1}, \sigma_{c=1}) + \pi_{c=2}\phi(y; \mu_{c=2}, \sigma_{c=2}) \quad (2)$$

Où φ(.) est la densité de loi Normale, d'espérance μ_c et de variance σ_c, pour c = 1, 2.

Un exemple classique de mélange fini de lois normales s'observe avec l'analyse des déterminants du revenu sur le marché du travail dans les pays en développement,

susceptible de différer sensiblement entre les travailleurs formels et informels, deux segments a minima. La fonction de densité globale du revenu est un mélange de deux sous-fonctions, l'une pour les travailleurs formels et l'autre pour les travailleurs informels (lui-même souvent subdivisés en deux sous-segments selon l'idée de Perry et al. (2007) et Günther et Launov (2012)). En déterminant ces sous-groupes, il est alors possible d'étudier distinctement l'effet de certains facteurs sociodémographiques sur le revenu des travailleurs informels et formels (Salem et Bensidoun, 2012).

En définitive, un modèle FMM peut s'écrire sous la forme d'une somme de plusieurs distributions de lois normales :

$$f(y,\theta) = \sum_{c=1}^{C} \pi_c \phi(y;\mu_c,\sigma_c) \tag{3}$$

Pour un nombre de composantes C fixé, nous cherchons à estimer les paramètres inconnus du modèle. Une formulation analytique de ces paramètres n'étant pas disponible, il est nécessaire d'utiliser des méthodes itératives d'estimation.

Une fois l'estimation des paramètres obtenue, le théorème de Bayes permet de déduire la probabilité *ex post* pour une observation i d'appartenir à un sous-groupe c :

$$\pi_{ic} = \frac{\pi_c \phi(y_i;\mu_c,\sigma_c)}{\sum_{c=1}^{C} \pi_c \phi(y_i;\mu_c,\sigma_c)}$$

Ces probabilités individuelles peuvent être utilisées pour classer les observations dans les différents sous-groupes.

Mélange de modèles de régression linéaire (FMRM)

Dans la section précédente les régresseurs étaient introduits dans les probabilités du mélange. Désormais, ils sont introduits dans les FDP des sous-populations. Cette approche permet simultanément: d'identifier des sous-groupes homogènes au sein de la population et d'estimer pour chacun d'eux un modèle de régression spécifique. Nous ne présenterons ici que le cas d'un mélange de régressions linéaires dans la mesure où ce cas reste le plus utilisé en sciences sociales (dérivé de Deb et al., 2011; Salem et Bensidoun, 2012; Günther et Launov, 2012).

Étape 1 - Nombre de composantes. Avant d'entamer la procédure FMM, il est nécessaire de vérifier que l'échantillon contient un nombre suffisant d'observations afin de supporter le partitionnement, différents travaux suggèrent un minimum de 30 individus par groupe (Garver et al., 2008). De plus, nous devons partir du postulat que le nombre de groupes est inférieur au nombre d'observations, ce qui revient à capter une "hétérogénéité inobservée entre les groupes" plutôt qu'entre les individus (Salem et Bensidoun, 2012: 2).

Par définition, lorsque l'on applique un modèle FMM, le nombre de groupes n'est pas connu a priori, nous devons alors l'inférer de nos données. Le choix du nombre de composantes C peut alors être effectué avec différentes méthodes, selon que l'on place un intérêt plus grand dans la qualité de l'ajustement de la densité ou dans la détection de sous-groupes distincts (McLachlan et Peel, 2000; Peel et McLachlan, 2000). La grande

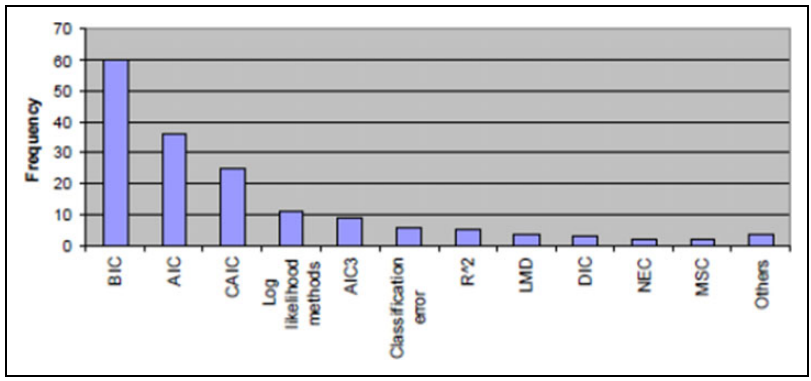

Figure 2. Métaanalyse des méthodes utilisées dans 108 articles en Marketing utilisant les FMM (Tuma et Decker, 2013)

majorité des méthodes consacrées à l'estimation des C groupes rassemblent des procédures basées sur le critère d'information (Figure 2). Ainsi, si l'intérêt principal concerne l'estimation d'une densité, une méthode adéquate consiste à sélectionner la valeur C qui minimise le critère AIC[5] (Akaike, 1974) *et/ou* le BIC[6] (Schwarz, 1978):

$$AIC = -2\hat{L} + 2K$$
$$BIC = -2\hat{L} + (3K - 1)\log(n)$$

où $\hat{L}$ est l'estimateur de la log-vraisemblance, n est le nombre d'observations et les facteurs 2K et 3K − 1 correspondent au nombre de paramètres à estimer. Une précision est utile : comme la log-vraisemblance augmente avec le nombre de paramètres, il faut utiliser un critère qui pénalise la vraisemblance afin de pouvoir comparer des modèles avec un nombre de paramètres différents. Ce critère fournit un estimateur convergent du paramètre K dans les modèles de mélanges (Keribin, 2000). En définitive, nous devons alors choisir le modèle qui minimise les valeurs de l'AIC et du BIC.

Étape 2 - Mélange de régression linéaire. Le modèle de régression linéaire simple consiste à spécifier l'espérance conditionnelle d'une variable dépendante y comme une fonction linéaire de plusieurs variables explicatives :

$$E(y|X) = x\beta$$

où x est une matrice de variables explicatives et β l'ensemble des paramètres.

"Une approche consiste à spécifier de la même manière les moyennes des fonctions de densité de chacun des sous-groupes du mélange. Si nous considérons un mélange fini de distributions Normales de variances σ 2 c (cf. p.4), nous pouvons réécrire l'équation (1) de manière à obtenir un mélange de modèle de régression linéaire" (Ahamada et Flachaire, 2008: 111) :

$$f(y|x;\theta) = \sum_{c=1}^{C} \pi_c \phi(y|x;\beta_c,\sigma_c) \quad (4)$$

Les moyennes des densités ne sont plus des paramètres directement estimés (comme lors l'équation 1), elles sont désormais conditionnelles aux valeurs des variables explicatives x, et obtenues à partir de l'estimation du modèle :

$$\mu_c = x\beta_c$$

Si l'on considère deux composantes (C = 2), cela revient à supposer le modèle suivant :

$$\begin{aligned} \textit{Groupe } 1 \ &: y = \ x\beta_1 + \varepsilon_1 \\ \textit{Groupe } 2 \ &: y = \ x\beta_2 + \varepsilon_2 \end{aligned}$$

Avec :

$$\begin{aligned} \textit{Pour le groupe } 1 \ &: \ \varepsilon_1 \sim N(0;\ \sigma_1^2) \\ \textit{Pour le groupe } 2 \ &: \ \varepsilon_2 \sim N(0;\ \sigma_2^2) \end{aligned}$$

où ε_1 et ε_2 sont des termes d'erreur indépendants et identiquement distribués, suivant une loi Normale et de variance respectives σ_1^2 et σ_2^2.

Ce modèle suppose que la population est composée de deux sous-groupes pour lesquels la relation entre la variable dépendante et les variables explicatives sont différentes. On retrouve ici l'idée générale proposée par la procédure de Gindling (1991) en ce qui concerne l'identification institutionnelle ou statistique des segments sur le marché du travail à San José (Costa Rica). Ce dernier applique successivement une correction du biais de sélection possible dans la probabilité d'appartenance aux différents groupes prédéterminés puis une régression linéaire sur chacun des segments en incluant l'inverse du ratio de Mills (λ) dans l'équation de détermination du revenu avant finalement de réaliser du test de Chow afin de tester l'hypothèse d'égalité des coefficients entre les différents groupes : $\widehat{\beta}_1 = \widehat{\beta}_c$ (Combarnous et Labazée, 2002).

Toutefois, la différence principale est que, dans le modèle de mélange, les observations sont issues de l'un des sous-groupes dans des proportions inconnues. Autrement dit, les sous-groupes ne sont jamais définis a priori, mais estimés de manière à être composés d'observations homogènes dans leur relation entre la variable expliquée y et les variables explicatives x.

Étape 3 - Algorithme EM et maximisation de la fonction de vraisemblance. Il existe une variété de procédures d'estimation de la fonction de vraisemblance dont les trois principales sont indiquées ici :

- Maximisation de la fonction de vraisemblance, à travers l'utilisation de méthodes itératives telles que l'algorithme *d'Expectation-Maximization*[7] (EM) ou de Newton-Raphson[8] (parfois les auteurs utilisent également l'algorithme de quasi-Newton ou le *Fisher's scoring*).
- Minimum χ 2
- Approches Bayésiennes (McLachlan et Peel, 2000; Frühwirth-Schnatter et al., 2012).

Nous ne retiendrons ici que la méthode d'estimation la plus employée: maximum de vraisemblance avec l'utilisation de l'algorithme EM. Développé initialement par

Dempster et al. (Dempster et al., 1977; Schafer, 1997), cet algorithme est une méthode itérative du calcul du maximum de vraisemblance. Il est basé sur le résultat central que le paramètre qui maximise la log-vraisemblance prédite augmente du même coup le logarithme de la vraisemblance observée. L'algorithme est donc la succession d'une étape (E), étape d'expectation, où on calcule l'espérance de la log-vraisemblance pour la valeur courante des paramètres puis d'une actualisation des paramètres en maximisant cette nouvelle fonction de ces mêmes paramètres, étape (M). L'algorithme converge sous des hypothèses de régularité vers un point stationnaire.

A chaque étape, l'algorithme EM augmente la valeur jusqu'à convergence vers un point stationnaire. Ces étapes sont répétées p fois jusqu'à ce que le critère de convergence soit satisfaisant.

Packages et commandes. De nombreux logiciels intègrent aujourd'hui des commandes permettant d'estimer des FMM ou des FMRM. Par exemple, R inclut les commandes Mclust et Fleximix. De même pour les versions de Stata, 11 et supérieur, les commandes fmm et fmmlc renvoient à l'estimation d'un modèle FMRM (Deb, 2012).

Sur R nous utilisons le package mclust pour estimer des mélanges de lois Normale. Pour l'estimation d'une densité bivariée :

```
install.package(mclust)
library(mclust) Melange = data.frame(x1, x2)
FMM <- Mclust(Melange)
surfacePlot(Melange, parameters=FMM$parameters)
```

Par ailleurs, le package fleximix permet d'estimer des mélanges de densité et de régressions. Pour un mélange de régressions linéaires avec c = 2 composantes, les variables explicatives étant x1 et x2 :

```
install.package(fleximix)
library(fleximix)
FMRM <- stepFleximix(y~x1+x2,k=2, nrep=4)
summary(FMRM)
```

Sur Stata la commande renvoyant à l'estimation d'un mélange de modèles de régression (mélange en deux composantes de lois Normales):

```
fmm y x1 x2, components(2) mix(normal)
```

Par ailleurs, la commande renvoyant les probabilités $\pi 1$ et $\pi 2$ ainsi que les critères AIC et BIC estimés pour le modèle :

```
fmm y x1 x2, components(2) mix(normal)
fmmlc, savec savep
```

Application – Déterminants du revenu dans l'économie formelle et informelle à Bogota

Informalité et marché du travail dans les PED. La prépondérance de l'économie informelle dans la grande majorité des pays en développement implique une analyse spécifique du marché du travail. Le travail informel qui s'opère soit au sein des unités de production informelles (employeur informel, employé informel, auto-emploi, etc.), soit dans le secteur privé formel ou dans le secteur public, rassemble un ensemble de pratiques et d'institutions sociales qui sont potentiellement différentes du travail formel (OIT, 2013).

Cette hétérogénéité peut se matérialiser sous des formes très diverses en fonction du contexte étudié. En effet, on observe par exemple des revenus plus faibles dans l'économie informelle, une absence totale ou quasi-totale de sécurité sociale ou professionnelle ainsi qu'une exacerbation des formes de discrimination de genre ou ethniques (Harriss-White, 2010). Ainsi, les travailleurs de l'économie informelle sont particulièrement vulnérables et font face à une précarité plus importante.

Partant de la constatation de l'existence de deux *sous-marchés* du travail ayant des dynamiques différentes, certains travaux empiriques se sont focalisés sur la démonstration de la segmentation de marché. En utilisant des fonctions de gains de type mincerienne, Gindling (1991) montre qu'il y a des logiques de construction du revenu différentes entre le secteur informel, formel privé et formel public au Costa Rica. Plus récemment, les études empiriques ont posé la question de l'homogénéité au sein du "secteur informel" dans les PED. Fields (2011) considère que le marché du travail informel est lui-même constitué de deux segments : un segment inférieur qui est composé de travailleurs exclus du marché formel et un segment supérieur constitué de travailleurs ayant choisi le marché informel car les opportunités de revenus y sont plus intéressantes (Maloney, 2004). En utilisant un modèle FMM, Günther et Launov (2012) montrent que le marché du travail informel en Côte d'Ivoire est composé à 45 pour cent d'emploi involontaire qui relève d'une logique d'exclusion et à 55 pour cent d'autres formes d'emploi qui relèveraient plutôt d'une stratégie d'opportunité.

Spécificités du marché du travail à Bogota. À l'instar de nombreux pays andins, la Colombie est marquée par une libéralisation continue de son marché du travail et présente un fort taux d'emploi informel, de très fortes inégalités de revenu et de patrimoine, un faible contrôle des relations individuelles de travail et de fortes restrictions des droits collectifs (Deguilhem et Frontenaud, 2016; Delmas et al., 2016). A Bogota plus spécifiquement, le recul de ces normes du travail génère une forte polarisation de la qualité des emplois (Combarnous et Deguilhem, 2016). Plus de 72 pour cent de la population de la ville participe au marché du travail en 2013 (SDP, 2013). Pourtant, l'occupation de la force de travail est très différenciée selon le genre, 64 pour cent des travailleurs occupés étant des hommes. De plus, la majorité des emplois sont formels, mais le taux d'informalité reste de 35,6 pour cent au sens de la dernière définition de l'OIT. Le secteur du commerce y est pré- dominant et la grande majorité des individus occupe un emploi de salarié du secteur privé (49 pour cent) ou de travailleur à compte propre (35 pour cent), l'emploi salarié public ne concernant que 4,5 pour cent de la population active occupée.

En 2013, Bogota compte plus de 7,6 millions d'habitants et représente ainsi près de 17 pour cent de la population colombienne à cette date, soit un accroissement de 87 pour cent depuis 1985 (SDP, 2013). En dépit d'un faible taux de fécondité et d'une réduction structurelle du taux d'urbanisation annuel, passant de 7 pour cent entre 1950 et 1955 à 1,36 pour cent de 2010 à 2015, Bogota reste marquée par la transition due aux migrations internes. Elle constitue un "pôle du système territorial" en accueillant les populations déplacées de force par le conflit interne au pays (Dureau et al., 2015: 35).

Dans les années 1970, face à l'expansion de l'urbanisation informelle et à l'accroissement des inégalités, le gouvernement a développé une méthode de stratification socioéconomique des espaces urbains puis ruraux, afin d'introduire un mécanisme de

financement croisé des services municipaux. Six groupes homogènes sont ainsi établis sur la base des zones cadastrales, en considérant l'aspect physique du bâti et un ensemble de critères *géoéconomiques*. Ces "îlots" de "voisinage homogène" offrent une approximation acceptable de la hiérarchie sociale. Les strates les plus défavorisées (strates 1, 2 et 3) représentent près de 90 pour cent de la population en 2013 et elles bénéficient de subventions couvrant de 10 à 40 pour cent du coût de leurs services municipaux. A l'inverse, les strates les plus avantagées (strates 5 et 6) paient un coût additionnel de 20 à 40 pour cent pour ces mêmes services. L'introduction de ce financement croisé a accru la logique insulaire du développement et la ségrégation résidentielle à Bogota (Dureau et al., 2015: 113-114).

Par ailleurs, comparativement aux autres métropoles andines, Bogota présente un faible taux de pauvreté absolue (17 pour cent en 2011). Toutefois, cette moyenne cache l'hétérogénéité des situations au sein la capitale. En effet, ce taux demeure élevé dans le Sud de la ville et au sein des zones urbaines les plus pauvres : 40 pour cent dans la strate 1 (SDP, 2013). Lors des dernières années, cette hétérogénéité des niveaux de vie a généré un accroissement important de la concentration du revenu des ménages, l'indice de Gini passant de 0,51 en 2008 à 0,61 en 2013 (SDP, 2013).

Données utilisées. Les données utilisées pour cette étude empirique sont issues de la Grande Enquête Intégrée des Ménages (GEIH) de 2013, représentative à l'échelle nationale et municipale, réalisée par le Département colombien de la statistique. L'échantillon est composé de 8855 individus ayant plus de 18 ans, se déclarant actif occupé (salarié ou indépendant) à Bogota et dont le revenu est supérieur à 0[9].

La distribution du revenu horaire (en pesos) constitue notre variable dépendante, elle est exprimée en logarithme népérien. Dans la base de données GEIH, nous avons également rassemblé un ensemble de caractéristiques socioéconomiques et démographiques utilisées comme variables explicatives : un proxy de l'expérience que nous calculons grâce à la différence entre l'âge d'un individu et la durée de scolarisation, le genre, si le niveau d'éducation primaire, secondaire ou supérieur correspond au niveau maximum obtenu par l'individu, la strate socioéconomique d'appartenance, le statut matrimonial, le nombre d'individus dans le ménage et le secteur d'activité.

Dans la suite de notre étude, nous utilisons un mélange fini de fonctions de gains mincériennes (Gindling, 1991; Combarnous et Labazée, 2002), nous permettant d'expliquer le revenu horaire par l'ensemble des variables indépendantes détaillées ci-dessus.

Résultats

Choix du meilleur modèle de partitionnement. Dans un premier temps, nous réalisons une régression en Moindres Carrés Ordinaires sur l'ensemble de l'échantillon, avant de traiter l'hétérogénéité du marché du travail par un modèle FMRM (tableau 1). Cette estimation en MCO décrit la situation d'un marché du travail compétitif ou non-segmenté. Sous cette hypothèse, les caractéristiques personnelles, notamment les dotations en capital humain permettent de déterminer le niveau de revenu horaire. Ainsi, en contrôlant certaines caractéristiques socioéconomiques et démographiques, deux individus ayant le même niveau de capital humain obtiendront le même revenu.

Tableau 1. Sélection du modèle AIC et BIC*

	AIC	BIC
1 segment (MCO)	18797.59	18967.56
2 segments (FMRM)	17418.24	17779.42
3 segments (FMRM)	16974.53	17519.84

*Nous avons utilisé le BIC ajusté par le nombre d'observations.
Source : Auteurs.

Dans un second temps, nous estimons un modèle FMRM en répétant les différentes étapes présentées précédemment.

Nous estimons d'abord le modèle avec plusieurs segments possibles sur le marché. Ensuite, nous choisissons le meilleur découpage à l'aide des critères d'information BIC et AIC, reposant sur une hypothèse de parcimonie des modèles et tendant à préférer les modèles les plus simples. A ce stade, il est intéressant de noter que la qualité de l'ajustement peut augmenter avec la complexification du modèle. Toutefois, l'ajout de nouveaux paramètres n'améliore le modèle que si le gain en qualité de l'ajustement permet de compenser la perte de parcimonie. La comparaison des AIC et BIC pour les différentes estimations montre que ces critères sont minimisés lors d'une partition en trois segments. Ainsi, nous observons que l'hypothèse d'homogénéité du marché du travail est invalidée, en d'autres termes, trois logiques distinctes coexistent sur le marché du travail de la capitale colombienne. Enfin, si les effets différentiés du rendement de la scolarisation et de l'expérience sont clairement identifiables, nous observons également que l'un des trois segments semble plus favorable que les deux autres.

Segmentation et effets différentiés. En premier lieu, nous observons que le revenu horaire moyen est beaucoup plus élevé dans le segment 3 que dans les segments 1 et 2 (tableau 2). A l'instar de nombreux travaux récents sur la segmentation du marché du travail dans les PED (Fields, 2011; Salem et Bensidoun, 2012; Günther et Launov, 2012) nous pouvons qualifier le segment 2 comme un segment inférieur, le segment 3 comme un segment supérieur et le segment 1 comme un segment intermédiaire. Dans le tableau 2, nous constatons que les segments comportent respectivement 30,41 pour cent, 54,21 pour cent et 15,38 pour cent de la population active et qu'ils se distinguent nettement par bien des aspects.

Par ailleurs, il est intéressant de noter que la partition du marché du travail ne coïncide pas avec la distinction communément admise : formel/informel. En effet, les segments inférieur et supérieur contiennent un peu moins de 30 pour cent de travailleurs d'informels tandis que le segment intermédiaire en contient près de 49 pour cent. Confirmant certains travaux sur la question (Combarnous et Deguilhem, 2016), nous observons que sur le marché du travail de Bogota la distinction formel/informel ne constitue pas le déterminant essentiel des différents segments, la réalité sociale au sein de chaque groupe apparait beaucoup plus nuancée.

De plus, nous constatons que contrairement aux segments inférieurs et intermédiaires, la différence de salaire entre le formel et l'informel est beaucoup plus faible dans le segment supérieur. Ainsi, une certaine logique de préférence du statut d'informel, en

Tableau 2. Statistiques descriptives pour l'économie formelle et informelle dans chaque segment (GEIH de 2013)

	Segment 1	Segment 2	Segment 3
% des actifs occupés	30,41%	54,21%	15,38%
% de travailleurs formels	51,07%	70,64%	70,66%
% de travailleurs informels	48,93%	29,36%	29,34%
Revenu horaire moyen (pesos)	6401,71	4153,32	12744,23
	(9398,19)	(3387,24)	(17559,14)
Revenu horaire moyen dans	4624,091	3752,146	12544,83
l'économie informelle (pesos)	(8354,295)	(2726,265)	(21034,96)
Revenu horaire moyen dans	8105,159	4320,079	12827,04
l'économie formelle (pesos)	(10010,71)	(3614,008)	(15905,9)

Note: Nous observons sur la Figure 2 que les méthodes les plus fréquemment utilisées dans la littérature en marketing usant des FMM restent l'AIC (Akaike, 1974) et le BIC (Schwarz, 1978). Ainsi, nous pouvons déterminer le nombre "optimal" de groupes permettant la meilleure application du modèle de mélange fini. Les écart-types sont indiqués entre parenthèses.
Source: Auteurs.

raison de meilleures opportunités (Maloney, 2004; Perry et al., 2007), semble ressortir de l'analyse du segment supérieur. Toutefois, ce phénomène apparait assez marginal à Bogota puisqu'il ne concerne qu'une faible part de la population informelle isolée dans le segment supérieur. Inversement, il semble bien que la contrainte et la mise au travail par nécessité prédomine dans le segment 1 et 2 (tableau 2).

Le tableau 3 reporte les coefficients des variables explicatives pour une régression linéaire en MCO et pour les différents segments du marché du travail tels que déterminés par le modèle FMRM. Nous observons que les segments se distinguent nettement par leur fonction de gains. Ainsi, ces trois segments constituent trois groupes non-concurrents sur le marché du travail, ayant chacun des logiques institutionnelles propres, en particulier en ce qui concerne les rendements des niveaux de capital humain. Plus précisément, nous constatons que l'effet du capital humain (notamment mesuré par le niveau d'éducation maximum obtenu) sur le revenu est supérieur dans le segment 3 qu'il ne l'est dans les deux autres segments. La rétribution d'un niveau d'éducation supérieur par rapport à un niveau d'éducation primaire (au maximum) est plus de deux fois plus élevée dans le segment 3 que dans le segment 1 et plus de quatre fois plus élevée que dans le segment 2.

Le genre de l'individu n'influence pas significativement les revenus dans les segments inférieur et intermédiaire (tableau 3). Toutefois dans le segment supérieur, être une femme induit un baisse de salaire notable par rapport aux hommes. Ceci est intéressant dans la mesure où il s'agit du segment le moins vulnérable su marché du travail, indiquant que le phénomène de "plafond de verre"[10] est une réalité sociale à Bogota.

Conclusion

Dans cet article, nous présentons ce qui distingue les modèles statistiques de mélange des méthodes de partitionnement couramment utilisées en sciences sociales pour rendre compte de l'hétérogénéité existante au sein d'une population. Après avoir détaillé les différentes étapes de ces outils afin de favoriser leur compréhension et leur usage, nous

Tableau 3. Estimations FMRM et MCO pour le logarithme népérien du revenu horaire (GEIH de 2013)

	FMRM			
Variable	Segment 1	Segment 2	Segment 3	MCO[b]
Expérience	-0,004***	0,0008	0,0104***	-0,0001
	(0,0013)	(0,0006)	(0,0015)	(0,0006)
Genre	0,0657	0,0047	-0,0975**	0,0120
	(0,040)	(0,0161)	(0,0417)	(0,0168)
Éducation sec.	0,1303***	0,0682***	0,2159***	0,1074***
	(0,0487)	(0,0191)	(0,0527)	(0,020)
Éducation sup.	0,3331***	0,1757***	0,7736***	0,3709***
	(0,0587)	(0,0270)	(0,0755)	(0,0252)
Strate 2	0,271***	0,0174	0,0343	0,1218***
	(0,0598)	(0,0228)	(0,0625)	(0,0237)
Strate 3	0,5715***	0,0669***	0,2904***	0,3354***
	(0,0639)	(0,0246)	(0,0736)	(0,0259)
Strate 4	0,7218***	1,354***	1,6929***	1,036***
	(0,0833)	(0,0536)	(0,0915)	(0,0418)
Marié	0,0987**	0,0642***	0,0486	0,0838
	(0,0443)	(0,0185)	(0,0478)	(0,0199)
Nombre de pers. dans le ménage	−0,02691**	−0,0010	−0,0335**	−0,0248***
	(0,0117)	(0,0049)	(0,013)	(0,0052)
σ_c	0,9012	0,2447	0,4033	
	(0,015)	(0,011)	(0,020)	
π_c[a]	0,4138	0,3678	0,2185	
	(0,0185)	(0,020)	(0,0166)	
Constante	7,7689***	7,922***	7,843***	7,915***
	(0,0976)	(0,0354)	(0,0921)	(0,0378)
Log likelihood		-8560,65		
Wald[2]		4915,50***		
Adjusted R^2				0,302
N		8796		8796

Notes: Les écart-types entre parenthèses sont robustes. Toutes les régressions incluent également les différents secteurs d'activité.

[a]π_c est la probabilité qu'une observation soit dans le segment c.

[b]Chaque prédicteur est non-corrélé significativement avec les autres prédicteurs (VIF<5).

*p < 0:1, **p < 0:05, ***p < 0:01.

Source: Auteurs.

fournissons une illustration empirique à travers la segmentation du marché du travail urbain à Bogota. Par ce biais, nous avons pu observer que les problèmes de sélection du nombre de groupes ont été facilités par l'emploi des critères statistiques d'information. Enfin, la synthèse des résultats et la précision des effets différenciés obtenus ont permis de faire apprécier l'intérêt de ces méthodes.

Cependant, plutôt que d'explorer toutes les possibilités offertes par ces modèles et leurs variantes, nous avons souhaité mettre ici en perspective ces outils en présentant certaines de leurs limites.

L'utilisateur de ces outils de modélisation statistique doit être précautionneux quant au nombre et au choix des variables utilisées dans la mesure où il existe un risque de dégénérescence du modèle. En effet, la log-vraisemblance d'un mélange de distributions gaussiennes avec des variances différentes n'est pas bornée, il n'existe donc pas de optimum global. Cela conduit l'utilisateur à devoir contraindre l'algorithme utilisé afin que les variances des composantes ne soient pas trop différentes. Ainsi, la maximisation sur cet espace contraint de paramètres va permettre d'obtenir une solution qui possède les propriétés désirées. Pour éviter tout problème d'*identifiabilité*, il faut également imposer la contrainte : $\mu_1 < \mu_2 < \ldots < \mu_C$ ou $\sigma_1 < \sigma_2 < \ldots < \sigma_C$ (Ahamada et Flachaire, 2008: 103). Enfin, la log-vraisemblance étant multimodale, l'algorithme peut converger vers un maximum local (d'autant plus fréquent que les deux composantes ne sont pas clairement séparées).

Pour conclure, les modèles FMM et FMRM se révèlent être des méthodes efficaces en apportant certaines solutions aux problèmes posés par les techniques de partitionnement issues de l'analyse exploratoire multidimensionnelle. Ainsi, comme nous avons pu le constater à Bogota, ces outils constituent une approche intermédiaire permettant de progresser dans la connaissance des phénomènes étudiés et ouvrent certaines pistes de réflexion pour leur utilisation dans le champ des sciences sociales. Toutefois, comme l'évoquent Le Guen et Jaffreux (1989: 94) : "il faut garder à l'esprit que tout traitement sur l'information [statistique] n'est ni unique, ni neutre [. . .]", et ces méthodes connaissent elles-aussi leurs limites.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Les données que nous utilisons sont issues de la *Gran Encuesta Integrada de Hogares* de 2013 (GEIH).
2. La définition de l'économie informelle retenue dans cet article est sur celle établie par l'OIT en 2003. Elle repose sur deux piliers qui sont les activités informelles d'une part et l'emploi informel d'autre part. Les activités informelles sont définies comme étant celles menées au sein de petites entreprises comptant moins de cinq employés, non enregistrées officiellement et ne tenant pas de comptabilité écrite. L'emploi informel est défini comme étant l'emploi sans contrat ou non protégé, au sein d'entreprises formelles ou informelles (Hussmanns, 2004).
3. Lebart et al. (2006) présentent notamment le critère de Kaiser, la méthode de Cattell et celle d'Anderson.
4. Pour plus de détails sur ces techniques voir Diday et Simmon (1976) et Lebart et al. (2006).
5. Akaike Information Criterion ou AIC.
6. Bayesian Information Criteria ou BIC.

7. L'algorithme de Dempster et al. (1977) reste le plus utilisé dans l'ensemble des travaux faisant usage de cette méthode (Ahamada et Flachaire, 2008).
8. Méthode itérative pour laquelle la convergence n'est pas assurée (Peel et McLachlan, 2000).
9. Le revenu des indépendants est approximé en calculant la somme des recettes indiquées pour leur activité moins la somme des charges déclarées.
10. Le "plafond de verre" signifie que toutes choses égales par ailleurs, une femme sera moins bien payé qu'un homme en haut de la distribution des salaires (OIT, 2010).

Références

Adams CP (2016) Finite Mixture Models with One Exclusion Restriction. *Econometrics Journal*, forthcoming.

Ahamada I et Flachaire E (2008) *Econométrie non-paramétrique*. Paris: Economica.

Akaike H (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19(6): 716-23.

Andrews RL, Brusco M, Currim IS et Davis B (2010) An Empirical Comparison of Methods for Clustering Problems - Are There Benefits from Having a Statistical Model? *Review of Marketing Science* 8(1): 1-32.

Boston TD (1990) Segmented Labor Markets - New Evidence from a Study of Four Race-Gender Groups. *Industrial & Labor Relations Review* 44(1): 99-115.

Combarnous F et Labazée P (2002) L'emploi en Côte d'Ivoire - Mobilisation du travail et production de rapports sociaux. Université Montesquieu Bordeaux IV, série de recherche.

Combarnous F et Deguilhem T (2016)Urban Labor Market Revisited - Why Quality of Employment Matters in Bogota. SSRN (January 2016), Working paper.

Conway KS et Deb P (2005) Is Prenatal Care Really Ineffective? Or, Is the "Devil" in the Distribution? *Journal of Health Economics* 24(3): 489-513.

Cowell FA et Flachaire E (2007) Income Distribution and Inequality Measurement-: The Problem of Extreme Values. *Journal of Econometrics* 141(2):1044-72.

Cunha F, Heckman JJ et Schennach SM (2010) Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica* 78(3): 883-931.

Davidson R et Flachaire E (2007) Asymptotic and Bootstrap Inference for Inequality and Poverty Measures. *Journal of Econometrics* 141(1): 141-66.

Deb P (2012) FMM: Stata module to estimate finite mixture models. *Stata Journal*, http://econpapers.repec.org/software/bocbocode/s456895.htm.

Deb P et Trivedi PK (1997) Demand for Medical Care by the Elderly - A Finite Mixture Approach. *Journal of Applied Econometrics* 12(3): 313-36.

Deb P et Trivedi PK (2002) The Structure of Demand for Health Care -: Latent Class versus Two-part Models. *Journal of Health Economics* 21(4): 601-25.

Deb P, Gallo WT, Ayyagari P, Fletcher JM et Sindelar JL (2011) The Effect of Job Loss on Overweight and Drinking. *Journal of Health Economics* 30(2): 317-27.

Deb P et Trivedi PK (2013) Finite Mixture for Panels with Fixed Effects. *Journal of Econometric Methods* 2(1): 35-51.

Deguilhem T et Frontenaud A (2016) Régimes de qualité de l'emploi et diversité des pays émergents. *Revue de la régulation*, forthcoming.

Delmas B, Deguilhem T et Vernot M (2016) Towards an Interdisciplinary Approach of Quality of Employment in Latin America - Evidence from Bogota. SSRN (May 2016), Working paper.

Dempster AP, Laird NM et Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1-38.

Diday E et Simon JC (1976) Clustering Analysis. In: Fu PKS (ed.) Digital Pattern Recognition. *Communication and Cybernetics* 10: 47-94.

Dureau F, Lulle T, Souchaud S et Contreras Y (2015) Mobilités et changement urbain - Bogota, Santiago et São Paulo. Rennes: Presses universitaires de Rennes.

Everitt BS, Landau S, Leese M et Stahl D (2011) *Cluster Analysis*. Chichester: Wiley, 5th edition.

Fields GS (2011) Labor Market Analysis for Developing Countries. *Labour Economics* 18(1): 16-22.

Frühwirth-Schnatter S, Pamminger C, Weber A et Winter-Ebmer R (2012) Labor Market Entry and Earnings Dynamics - Bayesian Inference Using Mixtures-of-experts Markov Chain Clustering. *Journal of Applied Econometrics* 27(7): 1116- 37.

Ghosal S et Van Der Vaart AW (2001) Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities. *Annals of Statistics* 29(5): 1233-63.

Gindling TH (1991) Labor Market Segmentation and the Determination of Wages in the Public, Private-formal, and Informal Sectors in San Jose, Costa Rica. *Economic Development and Cultural Change* 39(3): 5856-605.

Garver MS, Williams Z et Taylor GS (2008) Employing Latent Class Regression Analysis to Examine Logistics Theory: An Application of Truck Driver Retention. *Journal of Business Logistics* 29(2): 233-257.

Günther I et Launov A (2012) Informal Employment in Developing Countries - Opportunity or Last Resort? *Journal of Development Economics* 97(1): 88-98.

Harriss-White B (2010) Work and Wellbeing in Informal Economies - The Regulative Roles of Institutions of Identity and the State. *World Development* 38(2): 170-183.

Hudson K (2007) The New Labor Market Segmentation - Labor Market Dualism in the New Economy. *Social Science Research* 36(1): 286-312.

Hussmanns R (2004) *Defining and Measuring Informal Employment*. Geneva: ILO (February), Bureau of Statistics Paper.

Keane MP et Wolpin KI (1997) The Career Decisions of Young Men. *Journal of Political Economy* 105(3): 473-522.

Keribin C (2000) Consistent Estimation of the Order of Mixture Models. *Indian Journal of Statistics, Series A* 62(1): 49-66.

Le Guen M et Jaffeux C (1989). La conjonction analyse de données et statistique inférentielle pour conduire à une meilleure perception visuelle. *Revue de Statistique Appliquée* 37(3): 75-97.

Lebart L, Piron M et Morineau A (2006) *Statistique exploratoire multidimensionnelle - Visualisations et interférences en fouille de données*. Paris: Dunod.

Leontaridi M (1998) Segmented Labour Markets - Theory and Evidence. *Journal of Economic Surveys* 12(1): 103-09.

Li J, Ray S et Lindsay BG (2007) A Nonparametric Statistical Approach to Clustering via Mode Identification. *Journal of Machine Learning Research* 8(8): 1687-723.

Lubke GH et Muthen B (2005) Investigating Population Heterogeneity with Factor Mixture Models. *Psychological Methods* 10(1): 21-39.

Maloney WF (2004) Informality Revisited. *World Development* 32(7): 1159-78.

McLachlan G et Peel D (2000) Mixtures of Factor Analyzers. In: McLachlan G et Peel D *Finite Mixture Models*. Chichester: Wiley, 238-56.

OIT (2010) *Women in Labour Markets - Measuring Progress and Identifying Challenges*. Geneva: ILO Publications.

OIT (2013) *Statistics of Work, Employment and Labour Underutilization - 19th International Conference of Labour Statisticians*. Geneva: ILO Publications.

Peel D et McLachlan GJ (2000) Robust Mixture Modelling Using the T Distribution. *Statistics and Computing* 10(4): 339-48.

Perry G, Maloney WF, Arias O, Fajnzylber P, Mason AD et Saavedra-Chanduvi J (2007) Informality - Exit and Exclusion. Geneva: World Bank Publications.

Piore MJ (1983) Labor Market Segmentation - To What Paradigm Does It Belong? *American Economic Review* 73(2): 249-53.

Salem MB et Bensidoun I (2012) The Heterogeneity of Informal Employment and Segmentation in the Turkish Labour Market. *Journal of the Asia Pacific Economy* 17(4): 578-592.

Schafer JL (1997) Analysis of Incomplete Multivariate Data. Boca Raton Fl: CRC Press.

Schwarz G (1978) Estimating the Dimension of a Model. *Annals of Statistics* 6(2): 461-64.

Servicio Distrital de Planeación (SDP) (2013) Segregación socioeconomica en el espacio urbano de Bogota DC. Bogota: Secretaria Distrital de Planeación et Universidad Nacional de Colombia.

Stahl D et Sallis H (2012) Model-based Cluster Analysis. *Computational Statistics* 4(4): 341-58.

Tuma M et Decker R (2013) Finite Mixture Models in Market Segmentation - A Review and Suggestions for Best Practices. *Electronic Journal of Business Research Methods* 11(1): 2-15.

Wedel M et Kamakura WA (2000). Mixture Regression Models. In: Wedel M et Kamakura WA, *Market Segmentation - Conceptual and Methodological Foundations*. Heidelberg: Springer, International Series in Quantitative Marketing, 101-24.

Wede M, Kamakura WA et Böckenholt U (2000) Marketing Data, Models and Decisions. *International Journal of Research in Marketing* 17(2-3): 203-08.