

Méthode d'analyse régionale sur données individuelles : Introduction aux modèles multi-niveaux

Magalie Dinaucourt,

*INSEE, Direction Régionale du Languedoc-Roussillon
274 Allée Henri II de Montmorency - 34064 Montpellier Cedex 2*

Résumé

L'objectif de ces travaux est d'explorer des méthodes d'analyses statistiques pour mesurer et comparer l'influence du contexte local sur un phénomène. Par exemple, résider dans une région peut constituer un handicap ou au contraire un avantage pour l'insertion des individus sur le marché du travail en raison du contexte économique et social de la région. Plusieurs méthodes ont été présentées dans la littérature pour mesurer des effets régionaux : méthode par régression linéaire, méthode shift and share, présentées par H. JAYET (1993) ou encore étude de la corrélation des résidus. La plupart de ces méthodes utilisent des données agrégées au niveau géographique étudié et permettent d'apprécier les écarts régionaux en isolant les effets structurels. Ces modèles n'autorisent cependant pas à conclure sur les individus, c'est ce que D. COURGEAU (1997) appelle l'erreur écologique. Il est donc nécessaire de travailler sur des modèles au niveau individuel prenant en compte l'environnement dans lequel l'individu vit.

Les méthodes classiques pour prendre en compte la dimension géographique consistent soit à réaliser des modèles régionaux, soit à intégrer dans le modèle individuel une variable d'appartenance à la région. Cette dernière méthode suppose toutefois un certain nombre d'hypothèses, par exemple que l'influence soit la même pour tous les individus. D'autres variables agrégées au niveau géographique comme le taux de chômage peuvent également être intégrées dans le modèle (modèle contextuel). Spécifier dans un modèle logistique sur données individuelles des indicateurs agrégés au niveau spatial étudié, permet d'obtenir des effets contextuels.

Les modèles multi-niveaux, initialement développés en sciences sociales et plus particulièrement dans l'éducation ont été présentés pour prendre en compte une dimension contextuelle dans l'analyse individuelle, par exemple l'appartenance à une classe ou une école dans l'analyse des niveaux scolaires. Ils mesurent ainsi des « effets de grappes ». Ces modèles sont adaptés pour prendre en compte une dimension géographique dans un modèle estimé au niveau individuel, ce qui permet de repérer des effets régionaux.

Summary

The aim of this work is to explore methods of statistical analyses to measure and compare the influence of the context on an event. For example, to reside in an area can constitute an handicap or on the contrary an advantage for the insertion on the labour market because of the economic and social context of this area. Several methods have been explained in the literature to measure regional effects, namely method by linear regression, method shift and share, presented by H. JAYET (1993) or analyse of the correlation of the residuals. Most of methods use aggregate data at the geographical level studied and make it possible to appreciate the regional variations by isolating the structural effects. Anyway, it is impossible to conclude about the individuals with these models, it is what D.COURGEAU (1997) names the "ecological error". It's necessary to work on models at the individual level taking into account individual's own environment.

The traditional methods to take into account geographical dimension consist in either carrying out regional models, or integrate in the individual model a variable of membership into the area. However, this method supposes a certain number of assumptions, for example that the influence is the same for everyone. Other variables incorporated at the geographical level like the unemployment rate can be also integrated in the model (contextual model). To specify in a logistic model on individual data of the indicators incorporated at the studied space level, make it possible contextual effects obtention.

The multilevel models, initially developed in social sciences and more particularly in education were presented to take into account a contextual dimension in the individual analysis, for example membership of a class or a school in the analysis of the school levels. They have been developed to take hierarchical structure into account. These models are adapted to take into account a geographical dimension in a model estimated at the individual level, which makes it possible to locate regional effects.

Comment prendre en compte la dimension spatiale des phénomènes étudiés ?

L'objectif de cette étude est de comparer les méthodes statistiques existantes pour prendre en compte l'influence de l'environnement sur l'individu. L'importance de tenir compte simultanément de caractéristiques individuelles et agrégées dans l'explication de phénomènes a été exprimée par de nombreux auteurs ces vingt dernières années, initialement dans le domaine des sciences de l'éducation (mise en évidence d'effets « classes ») puis en démographie. En effet, l'existence de corrélation entre des données agrégées ne permet pas de conclure sur les individus, c'est ce que D.COURGEAU (1997) appelle l'erreur écologique. Toutefois, travailler sur des données individuelles sans prendre en compte l'effet de l'environnement sur l'individu (corrélation entre individus d'une même grappe, classe ou région...) conduit à l'erreur atomiste et ainsi à la non-validité des hypothèses du modèle à un seul niveau. Le but de cette étude est de montrer l'apport des modèles multi-niveaux dans le cadre de cette problématique.

Nous nous intéressons à la mesure d'un effet géographique à partir de données individuelles. Le niveau d'agrégation choisie est la région. Pour mesurer des effets régionaux, nous comparerons les méthodes classiques existantes puis introduirons les modèles multi-niveaux.

Dans le cadre de l'étude, la probabilité de perdre son emploi selon un certain nombre de critères, individuels ou collectifs est examinée. Les données sont issues des enquêtes « Emploi » de l'INSEE de 1996 à 2000 et portent sur les personnes en activité une année n et au chômage l'année $n+1$.

Le modèle de base est donc le modèle logit suivant :

$$\text{Log}\left(\frac{p_i}{1-p_i}\right) = \alpha + X_i\beta$$

où p_i est la probabilité de perdre son emploi. X_i , le vecteur des caractéristiques individuelles (sexe, diplôme, mobilité géographique, type de contrat, condition d'emploi, catégorie socioprofessionnelle, secteur d'activité, ancienneté dans l'entreprise).

A partir de ce modèle global, il est légitime de se demander si quelle que soit la région de résidence, les probabilités sont identiques et s'il existe des facteurs plus ou moins à risque selon les régions.

1 - Modèles régionaux

Une première méthode pour répondre à cette interrogation est de réaliser un modèle par zone géographique. Cette analyse permet de mesurer l'effet de chaque dimension explicative pour chacune des régions.

Pour la région j le modèle suivant est appliqué : $\text{Log}\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + X_i\beta_j$

La comparaison des coefficients ($\alpha_j + \beta_j$) permet de tirer des conclusions quant-à des différences régionales des effets sur chaque dimension explicative.

La réalisation de cette méthode sur les données de l'enquête Emploi ne permet cependant pas de conclure sur des effets régionaux. En effet, le faible nombre d'observations dans les régions conduit à de forts écarts-types des coefficients n'autorisant pas la conclusion de significativité des écarts entre coefficients. Toutefois, l'Alsace se distingue de la majorité des régions par un moindre risque de chômage pour un grand nombre de caractéristiques alors que les régions Provence-Alpes-Côte d'Azur-Corse, Ile-de-France et Languedoc-Roussillon, présentent des risques plus importants pour l'individu de référence. Les femmes sont plus à risque dans des régions Picardie, et moins en Poitou-Charentes.

En pratique cette méthode est peu utilisable pour plusieurs raisons. D'une part, il est nécessaire d'avoir à disposition un grand nombre de données pour chaque zone géographique pour tester l'hypothèse d'égalité des coefficients régionaux. Plus le nombre d'observations est faible, plus les écarts-types des coefficients sont forts, ce qui rend les comparaisons impossibles. D'autre part, elle suppose que chaque zone géographique soit représentée dans les données. Enfin, elle se révèle fastidieuse si le nombre de zones à étudier est important.

2 - Modèle avec indicatrices de régions

Pour éviter les problèmes que pose le faible nombre d'observations disponibles pour chaque région, il est possible de partir d'un modèle national et de le compléter en intégrant des variables géographiques, comme des indicatrices de régions. Utilisée dans un grand nombre d'études, cette méthode permet de mesurer des différences inter-régionales par le biais d'indicatrices de régions, variables résumant des effets locaux inobservables.

Le modèle s'écrit alors ainsi : $Log\left(\frac{p_i}{1-p_i}\right) = \alpha + X_i\beta + 1_R \delta_r$

Où δ_r est le paramètre régional et 1_R l'indicatrice régionale.

Toutes choses égales par ailleurs, un individu de la région k a $\delta_k - \delta_j$ fois plus de risques que l'événement se produise qu'un individu de la région j .

Ainsi, dans le cadre de notre exemple, seule l'Alsace se distingue avec un risque plus faible que dans la majorité des régions. Dans le Nord-Pas de Calais, les régions méditerranéennes, l'Île-de-France, le risque de perdre son emploi est plus important que dans les régions Bourgogne, Franche-Comté, Pays-de-la-Loire, Bretagne, Poitou-Charentes et Alsace.

Cette méthode présente l'avantage d'être simple d'utilisation. Les données disponibles au niveau individuel sont souvent issues d'enquêtes nationales, ce qui constitue peu d'individus au niveau régional pour pouvoir analyser des sous-groupes identiques. Elle permet de comparer des effets régionaux globaux. Toutefois, elle suppose l'hypothèse suivante : l'effet régional ainsi mesuré est identique toute chose égale par ailleurs, c'est-à-dire qu'il est le même pour tous les individus. Cette hypothèse d'additivité du modèle est très forte, puisque par exemple, dans le cas de notre étude, cela revient à dire que le sur-risque de perte d'emploi en Languedoc-Roussillon est identique chez les salariés en CDI et chez les titulaires de la fonction publique. Une façon de détourner l'additivité du modèle consisterait à croiser les variables géographiques et le statut de l'emploi par exemple si on suppose que l'effet régional n'est pas le même selon les statuts. Toutefois, ceci accroît considérablement le nombre de variables du modèle et les résultats sont rarement significatifs.

3 - Modèles contextuels

Dans les modèles contextuels classiques, l'influence du contexte sur l'individu est approchée par une variable commune au niveau agrégé, par exemple, le taux de chômage des régions, le PIB régional...

Le modèle logit s'écrit alors :

$$Log\left(\frac{p_i}{1-p_i}\right) = \alpha + X_i\beta + A_{R_i} \delta_{R_i}$$

où A_{R_i} est l'ensemble des variables au niveau agrégé.

Dans notre étude, la part des individus en emploi à durée limitée dans la population active occupée a été intégrée au modèle de base. Cette part est calculée par région et estimée à partir du Recensement de la Population de 1999.

Les paramètres estimés sont les mêmes que pour le modèle logit simple France entière, excepté pour la constante. L'estimation du paramètre pour le taux d'emplois dits « précaires » est significativement différente de 0 au seuil de 10% et est égale à 0.0247. La positivité du paramètre témoigne du fait que dans une région, plus le taux d'emplois précaires est important, plus les individus, quel que soit l'emploi occupé, ont de risques de perdre leurs emplois. Il y a donc un phénomène de contagion, puisque même les individus en emploi stable, s'ils sont dans une région à fort taux d'emplois précaires ont plus de risques de perdre leurs emplois. Ainsi, dans les régions où le marché du travail est fragile, c'est-à-dire où la part d'emplois à durée limitée est plus importante, cette fragilité se constatera sur l'ensemble des actifs occupés. Cet effet est un effet contextuel puisqu'il agit sur l'ensemble des individus. L'environnement joue donc un rôle sur le phénomène individuel étudié.

Le modèle classique dans lequel sont intégrées les indicatrices de région peut être désigné comme un modèle contextuel dans la mesure où prendre en compte la région de résidence de l'individu est considéré comme une intégration de l'environnement dans lequel l'individu vit dans le modèle à travers une variable communautaire.

4 - Modèles Multi-niveaux.

Les méthodes d'analyse multi-niveaux ont été développées, en particulier, par H. Goldstein (Londres) et J.J. Hox (Amsterdam) dans les sciences de l'éducation et appliquées plus généralement par la suite, notamment en démographie et dans le domaine sanitaire. Elles sont en particulier utilisées pour rechercher des corrélations entre, d'une part, des indicateurs individuels et, d'autre part, des variables socio-économiques prises en compte simultanément à plusieurs niveaux : individu, région, etc. Elles permettent ainsi d'étudier de quelle manière l'environnement socio-économique des individus influe sur les associations statistiques observées au niveau individuel.

Pour comprendre ces modèles, il est important de comprendre pourquoi ils ont été développés. Dans les années soixante-dix, une étude réalisée auprès d'élèves avança que des élèves apprenant la

lecture selon une méthode dite « formelle » progressaient plus vite que les autres. Cette étude utilisait les méthodes d'analyses de régression multiple classique, travaillant uniquement sur le niveau individuel et ignorant l'appartenance à une classe. Les résultats furent statistiquement significatifs. Cependant en 1981, Aitkin démontra que si la même étude est réalisée en tenant compte des groupes auxquels appartiennent les élèves (classes, même professeur), les différences jusqu'alors significatives disparaissent et les élèves ayant enseigné selon la méthode « formelle » ne se montrent pas différents des autres par rapport à leur progrès scolaire. Cette étude est le premier exemple le plus important d'analyse multi-niveaux. Elle a démontré l'existence d'un effet « Classe ». Suite à ces travaux et à leurs conclusions, les problèmes liés à la non-prise en compte de la structure hiérarchique des données ont bien été reconnus. Cependant faute d'outils puissants, ils demeuraient difficiles à résoudre. Aujourd'hui, il existe des outils adaptés pour répondre à ce type de problème (SAS, MLWIN...).

A - Le Modèle

Soit y_{ij} mesurant le fait de perdre son emploi pour un individu noté i résidant dans une région notée j .

y_{ij} prend la valeur 1 avec la probabilité p_{ij} et 0 avec la probabilité $(1-p_{ij})$.

Soit J le nombre total de régions j .

Dans le cas du modèle logit, La probabilité p_{ij} pour que la caractéristique à estimer y_{ij} soit égale à 1 s'écrit :

$$P(y_{ij} = 1 / x_{ij}) = \frac{1}{1 + e^{-x_{ij}\beta}}$$

Avec x_{ij} l'ensemble des variables.

Si on se limite au niveau individuel, le modèle national dit à un seul niveau est de la forme :

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \Pi_{ij} = X_{ij}\beta = a_0 + a_1x_{1ij} + e_{ij}$$

Le modèle à un niveau postule que les observations soient indépendantes. Ainsi, les termes d'erreur sont eux-mêmes indépendants avec une moyenne nulle et une variance σ^2 constante. Le postulat de la variance constante implique l'hypothèse que la relation soit la même à l'intérieur de chaque région, puisque les e_{ij} sont indépendants. **Ce deuxième postulat est remis en cause par le modèle multi-niveaux, puisqu'il suppose une corrélation entre les individus d'un même niveau agrégé.**

Si un modèle par région était réalisé, nous aurions J modèles de la forme suivante :

$$\text{Pour la région } j : \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \Pi_{ij} = X_{ij}\beta = a_{1j} + a_{2j}x_{2ij} + u_{ij}$$

où a_{1j} et a_{2j} sont les paramètres de la $j^{\text{ème}}$ région, u_{ij} étant le résidu aléatoire de moyenne nulle et de variance σ_{ej}^2 .

On pourrait estimer autant de paramètre a_{1j} et a_{2j} qu'il y a de régions mais la faiblesse du nombre de données rend difficile la comparaison entre régions des paramètres si les écarts-types sont trop élevés. Il en est de même si les observations ne représentent pas toutes les entités agrégées.

Une autre façon d'introduire les régions dans l'équation est de supposer que les paramètres a_{1j} et a_{2j} sont aléatoires et vont varier d'une région à une autre. C'est un modèle multi-niveaux particulier, où l'on suppose que l'effet aléatoire agrégé (région) joue sur la constante du modèle et sur le paramètre de l'explicative binaire X_{ij} .

Ainsi on pose :

$$a_{1j} = a_1 + u_{1j}$$

$$a_{2j} = a_2 + u_{2j}$$

a_1 et a_2 sont les paramètres moyens ajustés sur toutes les régions, u_{1j} et u_{2j} sont des variables

aléatoires de moyenne nulle dont on va estimer les variances et covariance.

$$\sum_{j=1}^J u_{1j} = 0, \sum_{j=1}^J u_{2j} = 0$$

Le modèle devient :

$$\text{Log}\left(\frac{p_{ij}}{1-p_{ij}}\right) = \Pi_{ij} = a_1 + u_{1j} + (a_2 + u_{2j})x_{ij} + e_{1ij} = a_1 + a_2 * x_{ij} + (u_{1j} + u_{2j}x_{ij} + e_{1ij})$$

Une partie est indépendante de la région ($a_1 + a_2 x_{ij}$), et une partie est aléatoire qui dépend à la fois de la région et de l'individu ($u_{1j} + u_{2j}x_{ij} + e_{1ij}$). La partie aléatoire n'est donc plus normalement distribuée. Les variances au niveau agrégé et au niveau individu sont supposées constantes et permettent de résumer les termes e_{0ij} , u_{0j} et u_{1j} .

$$\text{var}(e_{1ij}) = \sigma_1^2 = 1$$

$$\text{var}(u_{1j}) = \sigma_{u1}^2$$

$$\text{var}(u_{2j}) = \sigma_{u2}^2$$

$$\text{cov}(u_{1j}, u_{2j}) = \sigma_{u12}$$

La variance au niveau individuel est égale à 1, et l'on travaillera sur les variances et covariances au niveau agrégé.

La fonction logit de la probabilité que y_{ij} soit égale à 1 dans la région j pour les individus n'ayant pas la caractéristique x_{ij} s'écrit :

$$\Pi_{1j} = a_1 + u_{1j}, \text{ sa variance entre régions est égale à } \sigma_{u1}^2$$

La fonction logit de la probabilité que y_{ij} soit égale à 1 dans la région j pour les individus ayant la caractéristique x_{ij} s'écrit :

$$\Pi_{2j} = a_1 + a_2 + u_{1j} + u_{2j}, \text{ sa variance entre régions est égale à } \sigma_{u1}^2 + 2\sigma_{u12} + \sigma_{u2}^2$$

L'étude de ces variances, covariances et des corrélations permet de définir l'effet régional.

Ainsi, dans un modèle multi-niveaux aléatoire, trois paramètres de plus que dans une régression logistique simple sont à estimer: $\sigma_{u1}^2, \sigma_{u12}, \sigma_{u2}^2$. L'interprétation des coefficients fixes est identique à celle d'une régression logistique simple.

Les paramètres non aléatoires estimés avec un modèle multi-niveaux sont en général proches de ceux que l'on obtient avec un modèle logit simple.

Quand l'effet des aléas liés à la caractéristique est non nul au niveau régional, on observe une forte augmentation de la dispersion de ces paramètres : Doublement de l'écart-type. Dans le cas de notre exemple, lorsqu'un modèle multi-niveaux sur la constante et la variable explicative « emploi à durée limitée » est appliqué, la variance du coefficient de la variable explicative (0.062) a doublé par rapport à la variance de ce paramètre dans le modèle logit simple (0.034), ce qui laisse supposer qu'il existe bien des effets régionaux.

B - Résultats de l'étude

L'analyse multi-niveaux a été appliquée aux variables explicatives de l'étude, (sexe, diplôme, mobilité géographique, type de contrat, condition d'emploi, catégorie socioprofessionnelle, secteur d'activité, ancienneté dans l'entreprise). Des effets régionaux sur la perte d'emploi plus ou moins importants ont pu être mis en évidence :

Il existe un effet régional global, certaines régions telles que le Languedoc-Roussillon, la Provence-Alpes Côte d'Azur et le Nord présentant des sur-risques pour leurs habitants de perdre leurs emplois, alors que d'autres régions présentent de risques moindres : Alsace, Franche-Comté, Poitou-Charentes, Bretagne.

Pour les emplois dits « stables », il existe de fortes disparités régionales, qui s'atténuent pour les emplois à durée limitée. Dans les régions où le risque est plus fort qu'ailleurs, l'influence de la caractéristique « Emploi à durée limitée » est plus faible, en revanche pour les régions où le risque est faible, l'influence de cette caractéristique y est plus forte. Le Nord-Pas-de-Calais et la Picardie font exception : les risques de chômage sont encore plus importants qu'ailleurs pour les individus en emploi à durée limitée.

Pour les titulaires de la fonction publique, le risque de perte d'emploi est très faible quelle que soit la région. Toutes les régions ont un risque de même ordre. L'effet régional est donc nul pour cette sous-population. Le modèle vérifie notre hypothèse d'inégalité des effets régionaux selon le statut.

Pour les femmes, les disparités régionales même si elles existent sont atténuées comparativement aux disparités régionales sur le marché du travail masculin. Ce résultat peut s'expliquer par une moindre hétérogénéité du marché du travail féminin comparativement à celui des hommes. Des effets de structure non pris en compte dans le modèle seraient une cause des dispersions régionales plus importantes des hommes.

Dans les régions où la probabilité de perdre son emploi figure parmi les plus faibles, la probabilité des femmes est supérieure à celles des hommes. En revanche, dans les régions où la probabilité de perdre son emploi figure parmi les plus fortes, comme le Languedoc-Roussillon, la Provence-Alpes-Côte d'Azur-Corse, les femmes sont moins ou autant pénalisées que les hommes. Si pour toutes les régions, changer de région de résidence pour une femme est très pénalisant, pour certaines régions cela l'est encore plus, comme l'Aquitaine, la Basse-Normandie, Poitou-Charentes et le Limousin. En revanche pour d'autres, l'effet régional atténue le risque, sans le rendre nul, comme l'Alsace, la Champagne-Ardenne, la Picardie et la Haute-Normandie. En Ile-de-France, le risque y est plus faible qu'ailleurs.

Les résultats de cette étude permettent de conclure sur l'existence d'effets régionaux, mais surtout permet d'établir que les effets régionaux ne sont pas identiques sur toutes les sous-populations, remettant en cause l'hypothèse du modèle avec indicatrice de région.

C - Conclusion

Les analyses multi-niveaux permettent de prendre en compte des effets de classes (ou de grappes). Toutefois, D.Courgeau a montré que les risques d'inférence erronée semblent plus importants dans les modèles multi-niveaux, même lorsque la caractéristique omise est indépendante de celles introduites dans le modèle initial. Une bonne précaution est de faire intervenir dans le modèle le plus grand nombre de caractéristiques ayant un effet sur le phénomène afin d'éviter au maximum le risque de conclure à un effet d'agrégation alors qu'il n'existe pas. En spécifiant au maximum notre modèle (détailler les statuts, les CS...), les effets régionaux auraient peut-être pu être expliqués par des effets de structure.

Si l'effet propre du niveau agrégé n'est pas significatif (pas de corrélation entre individus d'un même niveau agrégé), le modèle devient un modèle contextuel classique, à savoir sans hiérarchisation. Autrement, si cet effet est significatif, (par exemple si appartenir à une région a une influence sur le phénomène observé, même après avoir éliminé les effets de contextes), les estimations du modèle contextuel classique sont biaisées, puisque l'hypothèse indispensable aux méthodes classiques d'analyse à un niveau d'indépendance des observations (i.e. des termes d'erreurs résiduels des équations de régression) n'est plus valide. L'approche à un niveau violant cette dernière conduit à une sous-estimation des écarts-types et à des interprétations biaisées (cf. Goldstein). Les modèles multi-niveaux sont des outils à utiliser avec précaution.

Bibliographie

- H.Jayet (1993), Analyse spatiale quantitative. Une introduction, Economica.
D.Pumain, Th Saint-Julien, A.Colin (1997), L'analyse spatiale, Cursus Géographie.
M.Marpsat et A.Trognon (1992), Présentation générale du modèle Logit, Actes des journées de méthodologie statistique, INSEE Méthodes N°46-47-48.
FUJITA Masahisa, THISSE Jean-François, Economie géographique, Problèmes anciens et nouvelles perspectives, Annales d'économie et de statistique n°45, p37-87.
Harvey Goldstein, (1995), Kendall's library of statistics 3, Multilevel statistical models.
Daniel Courgeau, Brigitte Baccaïni (1997), Analyse multi-niveaux en sciences sociales, Population 4
Daniel Courgeau (2000), Vers une analyse biographique multi-niveaux, VIIèmes Journées de méthodologie Statistique, Paris.
M.Aitkin and N.Longford (1981), Statistical modelling of data on teaching styles, Journal of the Royal Statistical society.
Laurence BLOCH, Marc-Antoine ESTRADE, 1998-1999, Les formes particulières d'emploi en France : un marchepied vers les emplois stables ? , Portrait social, France.
Yannick L'Horty (1999), L'emploi précaire en France, regards sur l'actualité N°249.
INSEE (2001), De plus en plus de passage vers un emploi stable », Insee Première N°769.
Claude Michel (1999), Un risque de perte d'emploi plus élevé en Languedoc-Roussillon, Repères Synthèse N°25.
Yvette Grelet, Bernard Hillau, Daniel Martinelli (1994), Portraits régionaux de l'emploi et de l'insertion des jeunes, Série observatoire, CEREQ, Document N°95.