Population Analysis of Disabled Children by Departments in France

Diah Meidatuzzahra¹, Heri Kuswanto¹, Nicolas Pech², Amélie Etchegaray³

¹Department of Statistics, Faculty of Mathematics and Sciences, Sepuluh Nopember Institute of Technology
²Department of Mathematics, Faculty of Sciences, Aix-Marseille University
³Centre inter-Regional d’Études, d’Action et d’Information (CREAI), Provence-Alpes-Côte d’Azur (PACA) et Corse

E-mail: meyda.azzahra@gmail.com, kuswanto.its@gmail.com, nicolas.pech@univ-amu.fr, amelie.etchegaray@creai-pacacorse.com

Abstract. In this study, a statistical analysis is performed by model the variations of the disabled about 0-19 years old population among French departments. The aim is to classify the departments according to their profile determinants (socioeconomic and behavioural profiles). The analysis is focused on two types of methods: principal component analysis (PCA) and multiple correspondences factorial analysis (MCA) to review which one is the best method for interpretation of the correlation between the determinants of disability (independent variable). The PCA is the best method for interpretation of the correlation between the determinants of disability (independent variable). The PCA reduces 14 determinants of disability to 4 axes, keeps 80% of total information, and classifies them into 7 classes. The MCA reduces the determinants to 3 axes, retains only 30% of information, and classifies them into 4 classes.

1. Introduction

The term “disability” is defined as a limitation of a person's ability to interact with their environment, due to a permanent disability or non-permanent that leads to stress and moral disorder, intellectual, physical or social. Disability is the consequence of an impairment that may be physical, cognitive, mental, sensory, emotional, developmental, or some combination of these. A disability may be present from birth, or occur during a person's lifetime. Disability comes in multiple forms and ambiguous definition. It is to be distinguished from the disease or the accident, which can be the disability origin.

In France, the definition disability is governed by French law dated 11 February 2005 on the opportunities and the same rights, and the participation and citizenship of disability people, that “Disability of activity limitations or restrictions on participation in the social life suffered by a person, due to substantial continuing modification of one or more functions”[1].

One of the institutions cared for the disabled is CREAI branch PACA et Corse. CREAI collaborated with Population Environment Development Laboratory (LPED) Aix-Marseille University propounds a project "Geography of Disability". The purpose is to estimate of the population and establishment medical social service. Therefore, statistical analysis is required to estimate the disabled population of multiple databases which were defined as disability determinants.
In this study, statistical analysis performed to model the variations of disabled children (aged 0-19 years) population among French department consisting of 14 determinant variables of children disabilities based on six categories, namely, the professional category of social (CSP) of their parents, the level of education of their parents, the premature rates, the tax of revenues, alcohol consumption, facilities and services of medical social for disabled. The aim is to classify departments according to their profiles determinants (socioeconomic and behavioral profiles).

In [2] has used PCA to clustering application of elderly people autonomy-disability). However, in this paper applying two types of methods: principal components analysis (PCA) and multiple correspondence analysis (MCA) to review which one is the best methods for the interpretation of the correlation between the determinants of disability (independent variable). And then, hierarchical clustering can be used to classify the departments according to their profile determinants. The software to support the analysis is SAS (Statistical Analysis System).

2. Preliminaries
This section, describe the definitions that will be used in this paper.

2.1. Centre inter-Régional d’Etudes, d’Action et d’Information (CREAI)
CREAI PACA et Corse or Central interregional of studies, action and information branch Provence-Alpes-Côte d'Azur (PACA) and Corse in France was founded in 1965 and belongs to the National Association of CREAI (ANCREAI) for a person with the condition of vulnerability. The CREAI is a private organizations and non-profit status established by statute law in 1901 which is subsidized by the State to optimize information sharing, collaboration and develop synergies of the technical experts to reflection and observation in the sectors of social action and medico-social [3].

2.2. The Determinants of Disability
The hypothesis of the project "Geography of Disability" is the distribution unusual of the disabled population in the region. The distribution of this population related by multiple factors, for example, economic factors, education, environment, lifestyle, etc., called the determinants of disability. [4] [5] [6]. Six groups of determinants identified:

2.2.1. The professional category of social (CSP). The CSP based on the data of INSEE. Labor force of 15 years and more having a job by gender, age, and the CSP are divided into 6 positions, is show on table 1.

| Number | The professional category of social |
|--------|-----------------------------------|
| 1.     | Farmer                            |
| 2.     | Artisan, craftsman and trader     |
| 3.     | Manager and high professions      |
| 4.     | Intermediate professions          |
| 5.     | Employee                          |
| 6.     | Labor                             |

2.2.2. The education level of their parents. In [7] identified through HSM survey, 33% of the disabled population aged 20-59 years is not graduation (without diplome). The database is used BTX_TD_FOR2_2012 of INSEE. It is classify 4 of education level, is show on table 2.
Table 2. The classify of education level

| Number | The level of education                        |
|--------|----------------------------------------------|
| 1.     | Without diplome                             |
| 2.     | BEPC / BEP / CAP (Diplome or certificate of professional) |
| 3.     | BAC (Baccalaureat)                         |
| 4.     | BAC*                                        |

2.2.3. The Premature Rates. The [8] study showed the importance of disabling sequelae preterm infants, before 33 weeks of amenorrhea (WA), and among those born between 33 and 36 WA of age. According [8], if the preterm birth is increased, the risk of disability is also. The 8th day certificate (Cs8) data used as database from 2010 to 2012.

2.2.4. The Tax of Revenues. The report of revenues tax is derived from local INSEE Social and Tax File (Philosophy) data. The first quartile of income report is the average wage in the department below which is 25% of wages [3].

2.2.5. The Consumption of Alcohol. Consumption of alcohol among women is unknown, therefore the number of premature deaths due to overdose of alcohol (it cause alcoholic psychoses and alcoholic cirrhosis of the liver) in women under the 65 years old were selected.

2.2.6. The facilities and services of medical social for disabled. The facilities and services of medical social for 1000 children disabled from age 0-19 years old. According the CREAI data (2016), Paris and PACA region are the least equipped regions mainly located (below 8 places for 1000 children). The Lozère is the most departments equipped in medico-social places with 22 places for 1000 children and also the Orne (18 places for 1000 children) and the Creuse (17 places for 1000 children).

2.3. Principal Component Analysis (PCA)
A principal component analysis (PCA) is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables [9]. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Its general objectives are reduction and interpretation of dimension (axes) data without reducing significantly the characteristics of the data. PCA is also often used to avoid problems of multicollinearity between independent variables in a multiple regression model.

The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables [10].

2.4. Multiple Correspondence Analysis
Multiple correspondence analysis (MCA) is the factorial method which adapted to tables. It is a set of individuals which described by several qualitative variables. It can be presented in many different ways. In France, following the work of L. Lebart, the most common is to focus on the similarities with
correspondence analysis, a method designed to study the relationship between two qualitative variables [11].

In the perspective of simultaneously processing quantitative and qualitative variables for the same individuals, which is one of the strengths of multiple factor analysis (MFA), it is important to focus on the similarities between principal component analysis (PCA) and MCA [12].

MCA can also be viewed as a PCA applied to the complete disjunctive table (CDT). To do this, the CDT must be transformed as follows. Let $y_{ik}$ denote the general term of the CDT. $y_{ik}$ is equal to 1 if individual $i$ possesses the category $k$ and 0 if not. Let’s denote $p_k$, the proportion of individuals possessing the category $k$ [13].

2.5. Hierarchical Ascendant Classification (HAC)

A hierarchical ascendant classification or hierarchical clustering method is a procedure which represents the data as a nested sequence of partitions. An example of the corresponding graphical representation, called a dendrogram. It is important to note that the height of a node is proportional to the distance between groups it links. Consequently, the shape of a dendrogram gives information on the number of clusters in a data-set. Thus, cutting a dendrogram horizontally engineers a clustering. The number of clusters is selected by three criteria. The diagram below represents the Cubic Clustering Criterion (CCC), pseudo F and pseudo t2 as a function of the number of clusters.

3. Methodology

In this section, explain of the steps in this research.

3.1. Data Collection

In the data collection is the first step in a study. The data used in this research is secondary data obtained from [15] [3] and the amount of department in France which is 96 departments as an observation.

3.2. Initial Processing

Initial processing stages to prepare the determinant of disability data. There are 14 determinant variables of children disabilities that will be used to process data, is show on table 3.
Table 3. The total of determinants variables

| Number | The determinant variables |
|--------|---------------------------|
| 1.     | Farmer                    |
| 2.     | Artisan, craftsman and trader |
| 3.     | Manager and high professions |
| 4.     | Intermediate professions  |
| 5.     | Employee                  |
| 6.     | Labor                     |
| 7.     | Without diplome           |
| 8.     | BEPC / BEP / CAP (Diplome or certificate of professional) |
| 9.     | BAC (Baccalaureat)        |
| 10.    | BAC¹                      |
| 11.    | The Premature Rates       |
| 12.    | The Tax of Revenues       |
| 13.    | The Consumption of Alcohol|
| 14.    | The facilities and services of medical social for disabled |

3.3. Experiment and Data Processing
In this step will discuss the data processing for the PCA, MCA, and HAC method using software SAS. For each method will be discussed as follows:

For processing the data using PCA method:
1. Obtained the correlation matrix from determinant variables
2. Obtained the eigenvalue of the correlation matrix
3. Consider the number of principal components (axes)
4. Interpretation of the sample principal components

For processing the data using MCA method:
1. Construction of table the data
2. Construction of the complete disjunctive table
3. Consider the contribution of variables to the construction of axes
4. Interpretation the contribution of variables and departments

For processing the data using HAC method:
1. Selection the number of class
2. Characterization and location of class

4. Discussion and Result

4.1. Principal Component Analysis (PCA)
The PCA reduces 14 determinants of disability to 4 axes, keeps 80% of total information. PCA is also calculating the coordinates of the department and the contribution of the determinants variables per axis. It is not balanced because some departments have too strong contribution. Figure 1 describe the number of axes will be interpretation, and table 4 describe the eigenvalues of the correlation matrix.
4.2. *Multiple Correspondence Analysis*
The MCA reduces the determinants to 3 axes, retains only 30% of total information. However, the determinant of the alcohol consumption does not contribute to these axes. The MCA graph is difficult to interpret. The 56 variables supplementary and 96 departments do not identify the group of department.

4.3. *Hierarchical Ascendant Classification (HAC)*
The Hierarchical Ascendants Classification (HAC) issued by the ACP put the 96 departments into 7 classes. Class 1 is the biggest class because there are 35 departments. In class 2, there are 18 departments. Class 3 includes all the departments of the Mediterranean such as: Hautes Alpes, Alpes-de-Haute-Provence, Alpes-Maritimes, Vars, Bouches-du-Rhône, Vaucluse, Gard, Hérault, Aude, Pyrénées-Orientales Corsica (Corse-du-Sud and Haute-Corse). In class 4 are found 16 departments. Class 5 and Class 6 each only 6 departments. On the contrary, class 7 contains only two departments, Paris and The Hauts de Seine.

The HAC issued by the MCA put the 96 departments into 4 classes. In class 1, there are 23 departments and class 2 is the biggest class because there are 31 departments. Class 3 includes 23 departments and in class 4 only 19 departments.

5. Conclusions
Based on the results and the discussion on the methods, the PCA is the best method for interpretation of the correlation between the determinants of disability (independent variable). Although, a more balanced classification of the department is classification on MCA. The PCA reduces 14 determinants of disability to 4 axes, keeps 80% of total information. PCA is also calculating the coordinates of the department and the contribution of the determinants variables per axis. It is not balanced because some departments have too strong contribution. The Hierarchical Ascending Clusterification (HCA) issued by the PCA made it possible to group the departments into 7
clusters. With 35 departments, cluster 1 is the cluster with the most individuals. On the contrary, cluster 7 contains only two. The MCA reduces the determinants to 3 axes, retains only 30% of total information. However, the determinant of the alcohol consumption does not contribute to these axes. The HCA issued by the MCA put the 96 departments into 4 clusters. In cluster 1, there are 23 departments and cluster 2 is the biggest cluster because there are 31 departments. Cluster 3 includes 23 departments and in cluster 4 only 19 departments.

Acknowledgment
The author would like to thank the lectures who have provided advice in the preparation of this research and also thank to other parties who contributed to the completion of this analysis.

References
[1] OMS 2011 Organisation Mondiale de la Santé
[2] Combes C, Azema J 2010 Clustering Using Principal Component Analysis: Application of elderly people autonomy-disability
[3] Le Centre interregional s’Études, d’Action et d’Informations (CREAI) en faveur des personnes en situation de vulnérabilité
[4] Vigneron E. 2011 Inégalités de santé, inégalités de soins dans les territoires français : Etat des lieux et voies de progrès. Elsevier Masson, France 194p
[5] Salem G., Rican S., Kurzinger M.L. 2000 Atlas de la santé de France, volume 2 : comportements et maladies. John Libbey, 222p.
[6] Vigneron E. 2013 Inégalités de santé, inégalités de soins dans les territoires français : Les tribunes de la santé, France p41-53.
[7] Espagnacq M. 2015 Populations à risque de handicap et restrictions de participation sociale Dossiers solidarité et santé, Drees. n°68, 2015, 18p.
[8] Etude épidémiologique sur les petits ages gestationnels (EPIPAGE), epidemiologic study on small-for-gestational-age children, France
[9] Johnson R.A, Wichern D.W 2007 Applied Multivariate Statistical Analysis United States of America
[10] Baccini A. 2010 Statistique descriptive multidimensionnelle, Publications de l’institut de mathématiques de Toulouse
[11] Dagneli P. 1975 Analyse statistique à plusieurs variables, Les presses agronomiques de gembloux, 362p
[12] Tenenhaus M. 2010 Analyse Factorielle des Correspondance Multiple et Classification Ascendante Hierarchique, France
[13] Bressoux P. 2010 Modélisation statistique appliquée aux sciences sociales, De boeck, 464p.
[14] Ribert A., Ehanji A., Lecourtier Y. 1999 An Incremental Hierarchical Clustering Canada
[15] INSEE, The National Institute of Statistics and Economic Studies collects, analyses and disseminates information on the French economy and society