Syllabus Minor B 214 – Data mining

Module : 214	**Data mining**
UE coordinator	Séverine Deguen, PhD Professor of biostatistics, Department of EPI-BIOSTAT, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr
Dates	November 14th to 18th 2016
Credits/ECTS	3 ECTS
Duration	5 days of 6 hours = 30 hours
UE description	Data mining is a statistical method used in public health to analyse high-dimensional data sets. To analyse the quantitative or qualitative variables, or both types, different methods are available. Data lining techniques are particularly adequate to synthetizing variables highly correlated and eliminating colinearity problem in multiple regression, for example, to constructing composite index such a socioeconomic deprivation index. These techniques also enable the design of homogeneous groups of statistical units. They also can be used for preparing sampling procedures in epidemiological studies. Each day is designed to alternate between theory and practice.
Prerequisite	Advanced core in biostatistics
Course learning objectives	***Learning objectives****: at the end of the module, the students should be able to:* - To be familiar with the most common methods: principal component analysis, cluster analysis. - To use the statistical function implemented in STATA software - To interpret the results including the statistical tables, such as contribution of variable on each components, correlation, eigen values and also correlation circle and dendogram.
UE Structure (details of sessions title/spaeker/date/duration)	Day 1: I Introduction to multidimensional methods – computer Lab Principal Component Analysis, technics for quantitative variables Monday November 14th, 9:00 -12:00 and 1:00 – 4:00 pm Day 2: Principal Component Analysis – computer lab Tuesday November 15th, 9:00 -12:00 and 1:00 – 4:00 pm Day 3: Cluster analysis - computer lab Wednesday November 16th, 9:00 -12:00 and 1:00 – 4:00 pm Day 4: Cluster analysis - computer lab Thursday November 17th, 9:00 -12:00 and 1:00 – 4:00 pm Day 5: computer lab - exam Friday November 18th , 9:00 -12:00 and 1:00 – 4:00 pm
Course requirement	Students will gain experience in using modern techniques to analyze high-dimensional public health data sets. Therefore they will be encouraged to prepare lab sessions, and self-practice on STATA .
Grading and assessment	Individual exam – last day of the module

Location	George Sand EHESP Campus in Paris
Readings	None

Session #1	**Introduction to multidimensional methods**
Speakers	Lecturer: Séverine Deguen, professor of biostatistics Department of EPI-BIOSTAT, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr
Learning Objectives	*At the end of the session, the students should be able to:* - To provide a general view of the different multidimensional approaches - To use different multidimensional technics - To perform a principal component analysis and to interpret the results
Duration	12 hours
Training methods	Lecture: Introduction to multidimensional methods, Principal Component Analysis, technics for quantitative variables Conference 1: Illustration: A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. Conference 2: Illustration: Data Analysis Technique: a Tool for Cumulative Exposure Assessment Practice on STATA software

Session # 2	Cluster analysis
Speakers	**Lecturer**: Séverine Deguen, PhD Professor of biostatistics severine.deguen@ehesp.fr
Learning Objectives	*At the end of the session, the students should be able to:* - To define the principle of a cluster analysis - To conduct a cluster analysis and draw a dendogram - To interpret the dendogram and and select the appropriate number of classes - To give the main characteristics of each class
Duration	12 hours
Training methods	Lecture: Cluster analysis (general principle, dendogram representation and interpretation, main characteristics of classes) Conference 1: Illustration: 'synthetic homogeneous neighborhoods' using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France Conference 2: Perception of the air quality in Lyon metropolitaen area Practice on STATA software