Syllabus Minor B 214 – Data mining

Module : 214	**Data mining**
Coordinator	Séverine Deguen, PhD Professor of biostatistics, Department of EOHS, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr
Dates	November 20th to 24th 2017
ECTS	3 ECTS
Duration	5 days of 6 hours = 30 hours
Location	Room : 409, EHESP 20 Avenue George Sand 93210 LA PLAINE ST DENIS
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 interpret the results including the statistical tables, such as contribution of variable on each components, correlation, eigen values and also correlation circle and dendogram.
Structure (details of sessions title/spaeker/date/duration)	Day 1: I Introduction to multidimensional methods – computer Lab Principal Component Analysis, technics for quantitative variables Project presentation Monday November 20th, 10:00 -12:00 and 1:00 – 4:30 pm Day 2: Conferences dealing with public health researches using principal component analysis – computer lab Tuesday November 21st 9:00 -12:00 and 1:00 – 4:30 pm Day 3: Factorial analysis – computer lab Wednesday November 22nd 9:00 -12:00 and 1:00 – 4:30 pm Day 4: Cluster analysis - computer lab Thursday November 23rd, 9:00 -12:00 and 1:00 – 4:30 pm Day 5: computer lab Friday November 24th , 9:00 -12:00 and 1:00 – 4:00 pm
Resources	All readings and materials will be posted on REAL
Course requirement	Students will gain experience in using modern techniques to analyze high-dimensional public health data sets.
Grading and assessment	Project by group (1/3 of the final mark) + Individual exam-2 hours (2/3 of the final mark) Note also that students will complete a questionnaire that assesses their own and their teammates' contributions to group work. All team members will receive the same grade except if it is clear that a student has not

	participated effectively (attended and contributed to meetings; made timely, helpful contributions; been constructive, etc.). In that case, the student's grade will be lowered accordingly.
Course policy	**Attendance & punctuality** **Regular and punctual class attendance is a prerequisite for receiving credit in a course.** Students are expected to attend each class. Attendance will be taken at each class. The obligations of attendance and punctuality cover every aspect of the course: - lectures, conferences, group projects, assessments, examinations, as described in EHESP Academic Regulations http://mph.ehesp.fr EHESP Academic Regulation Article. 3). If students are not able to make it to class, they are required to send an email to the instructor and to the MPH program coordinating team explaining their absence prior to the scheduled class date. All supporting documents are provided to the end-of-year panel. . Students who miss class are responsible for content. Any student who misses a class has the responsibility for obtaining copies of notes, handouts and assignments. If additional assistance is still necessary, an appointment should be scheduled with the instructor. Class time is not to be used to go over material with students who have missed class. **Lateness**: Students who are more than 10 minutes late may be denied access to a class. Repeated late arrivals may be counted as absences (See http://mph.ehesp.fr EHESP Academic Regulation Article. 3 Attendance & Punctuality) **Maximum absences authorized & penalty otherwise** Above 20% of absences will be designated a fail for a given class. The students will be entitled to be reassessed in any failed component(s). If they undertake a reassessment or they retake a module this means that they cannot normally obtain more than the minimum pass mark (i.e. 10 out of 20) **Exceptional circumstances** Absence from any examination or test, or late submission of assignments due to illness, psychological problems, or exceptional personal reasons must be justified; otherwise, students will be penalized, as above mentioned. Students must directly notify their professor or the MPH academic secretariat before the exam or before the assignment deadline. Before accepting the student's justification, the professor or the MPH academic secretariat has the right to request either a certificate from the attending physician or from a psychologist, or from any other relevant person (See http://mph.ehesp.fr EHESP Academic Regulation Article 4 Examinations). **Courtesy**: All cell phones/pages MUST be turned off during class time. Students are required to conduct themselves according to professional standards, eating during class time is not permitted during class time, such as course or group work.
Valuing diversity	Diversity enriches learning. It requires an atmosphere of inclusion and tolerance, which oftentimes challenges our own closely-held ideas, as well as our personal comfort zones. The results, however, create a sense of community and promote excellence in the learning environment. This class will follow principles of inclusion, respect, tolerance, and acceptance that support the values of diversity. Diversity includes consideration of: (1) life experiences, including type, variety, uniqueness, duration, personal values, political viewpoints, and intensity; and (2) factors related to "diversity of presence," including, among others, age, economic circumstances, ethnic identification, family educational attainment, disability, gender, geographic origin, maturity, race, religion, sexual orientation and social position.
Course evaluation	EHESP requests that you complete a course evaluation at the end of the school year. Your responses will be anonymous, with feedback provided in the aggregate. Open-ended comments will be shared with instructors, but not identified with individual students. Your participation in course evaluation is an expectation, since providing constructive feedback is a professional obligation. Feedback is critical, moreover, to improving the quality of our courses, as well as for instructor assessment.

Session #1	**Introduction to multidimensional methods** **Principal component analysis**
Speakers	Lecturers: Séverine Deguen, professor of biostatistics

	Department of EPI-BIOSTAT, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr Wahida Kihal , researcher in spatial epidemiology, CNRS, Strasbourg
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	13 hours
Training methods	Lecture: Introduction to multidimensional methods, Principal Component Analysis, technics for quantitative variables Conference 1: A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. Conference 2: 'Perception of the air quality in Lyon metropolitaen area Practice on R software
Session #2	**Factorial analysis of multiple correspondence**
Speakers	Lecturer: Séverine Deguen, professor of biostatistics Department of EPI-BIOSTAT, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr Wahida Kihal , researcher in spatial epidemiology, CNRS, Strasbourg
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 factorial analysis of multiple correspondence and to interpret the results
Duration	10 hours
Training methods	Lecture: Introduction to Multiple Factor Analysis - Factorial analysis of multiple correspondences Conference 1: Illustration: A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. Practice on R software

Session # 3	Cluster analysis
Speakers	Lecturer: Séverine Deguen, professor of biostatistics Department of EPI-BIOSTAT, EHESP – Sorbonne Paris Cité severine.deguen@ehesp.fr Wahida Kihal , researcher in spatial epidemiology, CNRS, Strasbourg
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	9 hours
Training methods	Lecture: Cluster analysis (general principle, dendogram representation and interpretation, main characteristics of

	classes) Conference 1: Neighborhoods' using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France' Practice on R software