UBS - Université Bretagne Sud, Vannes, Morbihan
UFR Sciences et Sciences de l'Ingénieur
Département Mathématiques, Informatique, Statistique
Campus de Tohannic - BP 573 - F-56017 Vannes cedex

# Livret Pédagogique

# Master

# Mention : Mathématiques Appliquées, Statistique

# Parcours : Data Science et Modélisation Statistique (DSMS)

## OBJECTIFS

L'objectif du Master DSMS est de former des experts exerçant dans le domaine des statistiques et des sciences des données. Cette offre de formation orientée à la fois vers les entreprises et vers la recherche est particulièrement attractive, car elle permet de développer des compétences nécessaires à l'essor des technologies de l'information qui irriguent tous les secteurs de l'économie et qui intéressent les entreprises de toutes tailles. L'analyse et le traitement de ces données hétérogènes, complexes et massives font de plus en plus appel aux derniers développements des mathématiques appliquées, de l'informatique et de la statistique. C'est pourquoi l'étroite imbrication d'enseignements dans ces disciplines permet aux étudiants d'acquérir des compétences transversales qui sont cruciales et d'allier des capacités d'abstraction pour concevoir des modèles numériques appropriés et pour les mettre en œuvre au moyen de technologies logicielles avancées. La variété des domaines scientifiques, des approches technologiques et des entreprises engendre une grande diversité des métiers relevant du domaine des sciences des données. Le Master DSMS répond à cette diversité par une large palette de spécialités au sein d'un même ensemble offrant une identité unique et un socle commun de connaissances permettant des parcours pluridisciplinaires en lien avec les laboratoires de recherche et les entreprises.

La formation se déroule en $4$ semestres et s'appuie sur les enseignements de la Licence Sciences, Technologies, Santé (semestre 1 à semestre 6). Ces enseignements sont complétés chaque semestre par des enseignements de Sciences Humaines et Sociales (SHS), l'apprentissage par projets et des stages favorisant l'ouverture vers le monde extérieur, et donnant lieu à la délivrance de 30 ECTS par semestre.

L'objectif des projets et des stages tout au long du parcours est d'offrir une formation ouverte sur la recherche scientifique et sur les besoins des entreprises. Les années M1 et M2 permettront d'acquérir des connaissances de haut niveau dans les domaines connexes aux activités des laboratoires IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires), Lab-STICC (Laboratoire en Sciences et Techniques de l'Information, de la Communication et de la Connaissance) et LMBA (Laboratoire de Mathématiques de Bretagne Atlantique). Le Master DSMS intègre ainsi une initiation à la recherche au cœur des sciences des données, destinée à former des scientifiques capables de s'adapter à l'évolution rapide des technologies de l'information dans des environnements numériques complexes.

## SEMESTRE 1

**Unités d'Enseignement Obligatoire**

- STA2105 : Modèles Linéaires Généralisés (Estimations et Prédictions)
- STA2121 : Statistique Bayésienne et MCMC
- INF1612 : Systèmes d'information opérationnels : base de données
- STA2122 : Programmation et traitement statistique des données
- STA2120 : Séries chronologiques et prévisions

**Enseignement Complémentaire (UEC)**

- SUCG401
  - ANG2102 : Anglais
  - ECN2102 : Droit

## SEMESTRE 2

**Unités d'Enseignement Obligatoire**

- STA2215 : Machine Learning et Big Data
- INF2204 : Systèmes d'Information décisionnels et entrepôts de données
- STA2123 : Modèles de durées et Analyse de Survie
- MIS2251 : Projet Tutoré
- STA2124 : Optimisation statistique et Business Intelligence

**Unités d'Enseignement Complémentaire (UEC)**

- SUCG402
  - ANG2202 : Anglais
  - COM2202 : Techniques d'expression

## SEMESTRE 3

**Unités d'Enseignement Obligatoire**

- STA2326 : Modélisation de données complexes
- STA2328 : Intelligence Artificielle et Deep Learning
- STA2325 : Statistique spatiale et Systèmes d'Information Géographique (SIG)
- STA2329 : Challenge Kaggle & Big Data (Hadoop, Spark)
- STA2321 : Machines à Vecteur Support et méthodes à noyaux

**Unités d'Enseignement Complémentaire (UEC)**

- CON2302 : Conférences et mini-cours
- ANG2306 : Anglais

## SEMESTRE 4

- STA2324 (Unités d'Enseignement Obligatoire) : Stage long (10 semaines minimum)

## Contacts

**Scolarité**
Sandrine Steinmann (sandrine.steinmann@univ-ubs.fr)

**Directeur des Études :**

— Semestre 1 Master 1 DSMS : Professeur Ion Grama (ion.grama@univ-ubs.fr)

— Semestre 2 Master 1 DSMS : Professeur Gilles Durrieu (gilles.durrieu@univ-ubs.fr)

— Master 2 DSMS : Professeur François Septier (francois.septier@univ-ubs.fr)

— Responsable des conférences du Master 2 DSMS : Jean-Marie Tricot (jean-marie.tricot@univ-ubs.fr)

— Responsable des stages du Master 2 DSMS : Evans Gouno (evans.gouno@univ-ubs.fr)

**Site web :** https://www.univ-ubs.fr/master-statistique

# UNITÉS D'ENSEIGNEMENT DU SEMESTRE 1

# STA2105
# Modèles linéaires généralisés (estimations et prédictions)

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

L'objectif de ce cours et de montrer comment on peut généraliser le modèle linéaire dans des situations où il ne donne pas des résultats satisfaisants. Nous analysons en détails la régression logistique, les données de comptage et les tableaux de contingence.

## Contenu

- Modèle linéaire. Condition d'utilisation. Les types des variables. Exemples.
- Famille exponentielle et modèles linéaires généralisés. Information de Fisher. Exemples.
- Estimation dans les modèles linéaires généralisés. Exemples.
- Inférence statistique pour les modèles linéaires généralisés. EMV et sa loi limite. Déviance. Exemples.
- Réponses binaires et régression logistique. Exemples.
- Régression logistique nominale et ordinale. Exemples.
- Données de comptage et modèle log - linéaire. Exemples.
- Tables de contingence. Exemples.
- Réduction de la dimension de l'espace des variables explicatives. Exemples.

## Prérequis

Probabilités MTH1303. Statistique mathématique STA1512.

## Bibliographie


- A. Doobson. An introduction to generalised linear models. Chapman and Hall 2002.
- P. McCullagh and J.A. Nelder. Generalized linear models. Chapman and Hall 1989.
- A. Antoniadis, J. Berruyer, R. Carmona. Régression non linéaire et applications. Economica 1992.
- P.A. Cornillon, E.Matzner-Lober. Régression. Théorie et applications. Springer 2005.


Fiche mise à jour le 08/10/2020

# STA2121
# Statistique Bayésienne et MCMC

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

Découvrir les bases de l'approche bayésienne des problèmes statistiques et s'initier aux outils de l'analyse bayésienne.

## Contenu

- Introduction : Généralités : Fréquentistes / Bayésiens
- Eléments de Théorie de la Décision - Modèle de décision - Règles de décision - Relation de préférence - Fonction de coût - Fonction de risque - Optimalité : minimaxité et admissibilité.
- Analyse Bayésienne - Lois a priori - Lois conjuguées - Lois non informatives - Estimateur de Bayes - Tests et régions de confiance.

## Prérequis

Statistique inférentielle, régression

## Bibliographie


- Lehmann E. L., Theory of Point Estimation, Wiley, 1983.
- Robert C., L'analyse statistique bayésienne, Economica, 1992.


Fiche mise à jour le 08/10/2020

# INF1612
# Systèmes d'information opérationnels : bases de données

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

Dans le cadre de la conception de systèmes d'information, l'étudiant sera capable d'intervenir sur les différentes étapes du projet, depuis la re-documentation du cahier des charges jusqu'à la mise en production des bases de données, quelles qu'elles soient.

## Contenu

- Modèles conceptuels des systèmes d'information
  - Merise
  - UML (cas d'utilisation, diagramme de classes et des séquences, diagramme d'activité, profiles)
  - modèle logique, modèle physique
- Modèle logique, modèle physique
- Programmation avancée des bases de données
  - PL/SQL ou Transact SQL
- Performance d'accès aux BD
  - indexation
  - optimisation de requêtes
  - tuning de bases de données
  - répartition

## Prérequis

Bases de données relationnelles et langages associés.

## Bibliographie


- Merise et UML pour la modélisation des systèmes d'information, Joseph Gabay. Dunod, 2004
- ORACLE 10g, guide du DBA, Kevin Loney, Bob Bryla. Oracle Press. 2005.
- Oracle Performance Tuning for 10g, Gavin Powell. Elsevier, 2005
- Oracle 10g, optimisation d'une base de données, Claire Noiraud, ENI, 2006
- UML2 pour l'analyse d'un système d'information, Chantal Morley, Jean Hugues, Bernard Leblanc. Dunod, 2006.
- Microsoft SQL Server 2005, guide de l'administrateur, William Stanck, Microsoft Press, 2006
- SQL Server 2008, SQL, Transact SQL, Jérôme Gabillaud, ENI, 2008


Fiche mise à jour le 08/10/2020

# STA2122
# Programmation et traitement statistique des données

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

L'objectif de ce cours est de permettre aux étudiants de maîtriser les concepts de programmation R et SAS tout en approfondissant un ensemble de techniques statistiques.

## Contenu

1. Introduction aux logiciels R, Python et SAS
2. Programmation statistique sous R, Python et SAS
3. La proc SQL (SAS)
4. Le langage MACRO (SAS)
5. SAS IML Studio (SAS)
6. Simulation, modélisation et analyse de données sous R, Python et SAS
7. Exemples d'applications sur des données réelles (R, Python et SAS)

## Prérequis

Eléments de programmation, Algorithmique.

## Bibliographie


— H. Kontchou-Kouomegni, O. Decourt. SAS : Maîtriser SAS Base et SAS Macro, Dunod, 2006
— S. Ringuedé. SAS : Introduction au décisionnel - Méthode et maîtrise du langage, Pearson Education, 2008
— E. Duguet. Introduction à SAS, Economica, 2004
— F. Husson, S. Lê, J. Pagès (2009) Analyse de données avec R, Presse Universitaires de Rennes
— G. Sawitzki (2009) Computational Statistics : an introduction to R ; Chapman & Hall


Fiche mise à jour le 08/10/2020

# STA2120
# Séries chronologiques et prévisions

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

Les techniques de décomposition des séries temporelles couramment utilisées dans la prévision des ventes, par exemple, et l'utilisation des processus stochastiques pour le traitement des séries chronologiques et, en particulier, les séries rencontrées dans le domaine de la Finance.

## Contenu

- Analyse spectrale
- Modèles exponentiels de Holt-Winter
- Les modèles de Box & Jenkins
- Prévisions
- Modèles d'intervention ARMAX
- Modèles ARCH et GARCH
- Filtrage linéaire (filtre de Kalman)

## Prérequis

Statistique inférentielle, régression

## Bibliographie


- Bourbonnais, R. Terraza, M. (2004), Analyse des séries Temporelles, Dunod, Paris.
- Gouriéroux, C. Monfort, A. (1995), Séries Temporelles et Modèles Dynamiques,
- Tenenhaus, M. (2007), Statistique ; Méthodes pour décrire, expliquer et prévoir, Dunod, Paris.


Fiche mise à jour le 08/10/2020

# Projets courts

## Modalités pédagogiques

Projets réalisés pendant les UE(s)

## Objectifs

Donner aux étudiants une ouverture à la recherche et au travail collaboratif.

## Contenu

Plusieurs projets courts seront données aux étudiants au sein des UE.

## Prérequis

UE(s) enseignés en semestre 7 ou semestres précédents.

## Bibliographie

Fiche mise à jour le 08/10/2020

# SUCG401
# Enseignement complémentaire

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

- **Droit du travail** : Expliquer la situation du salarié dans l'entreprise et celle de l'employeur en apportant le cadre juridique de la relation : droits et obligations des uns et des autres, la représentation des salariés dans l'entreprise - le déroulement du contrat de travail.
- **Droit de l'information** : Donner les bases du droit de l'information et des créations informatiques.
- **Anglais**

## Contenu

- **Droit du travail**
  - L'environnement juridique du droit du travail (sources et structures)
  - Le contrat de travail (les opérations d'embauche, les caractéristiques spécifiques, le déroulement du contrat - durée, rémunération)
  - La rupture du contrat de travail
  - La représentation des salariés dans l'entreprise
- **Droit de l'information**
  - Créations informatiques et acteurs
  - Montages contractuels et responsabilités
  - Montages contractuels spécifiques
  - Les licences logicielles
  - Création administration de sites web

## Prérequis

Aucun

## Bibliographie

Code du travail

Fiche mise à jour le 08/10/2020

# ECN2102
# Droit

## Modalités pédagogiques

Matière de l'UE d'Enseignement Complémentaire / Culture Générale. TD (24h) en présentiel.

## Objectifs

Droit du travail Expliquer la situation du salarié dans l'entreprise et celle de l'employeur en apportant le cadre juridique de la relation : droits et obligations des uns et des autres, la représentation des salariés dans l'entreprise - le déroulement du contrat de travail.
Droit de l'information Donner les bases du droit de l'information et des créations informatiques.

## Contenu

- Droit du travail
  - L'environnement juridique du droit du travail (sources et structures)
  - Le contrat de travail (les opérations d'embauche, les caractéristiques spécifiques, le déroulement du contrat - durée, rémunération)
  - La rupture du contrat de travail
  - La représentation des salariés dans l'entreprise
- Droit de l'information
  - Créations informatiques et acteurs
  - Montages contractuels et responsabilités
  - Montages contractuels spécifiques
  - Les licences logicielles
  - Création administration de sites web

## Prérequis

Aucun.

## Bibliographie

Code du travail.

Fiche mise à jour le 08/10/2020

# UNITÉS D'ENSEIGNEMENT DU SEMESTRE 2

# STA2215
# Machine learning et Big Data

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

- comprendre les enjeux de l'approche statistique de l'apprentissage,
- acquérir les compétences pour analyser de manière efficace des données de grande taille,
- s'initier aux outils spécifiques : Hadoop MapReduce et Spark.

## Contenu

- Première partie :
  - Introduction : généralités sur l'apprentissage statistique.
  - Régression
  - Quelques machines pour la classification :
    - la régression logistique,
    - l'analyse discriminante linéaire,
    - k–plus proche voisin.
  - Évaluation des machines.
  - Arbres de décisions : bagging, boosting, forêts aléatoires.
- Seconde partie :
  1. Linux, 4V of big data, Evaluation method for this course, bottleneck with typical system for big data, parallel computing, limitations and solutions.
  2. Hadoop Mapreduce.
  3. Descente de gradient stochastique
  4. Dplyr, Spark.
  5. Big data and Python, Eléments de cloud computing.
- Outils : R, RStudio, Linux, Spark, Hadoop.

## Prérequis

- Algèbre/Analyse/Statistique,
- Connaissances basiques en apprentissage machine (régression linéaire, classification), • Bonne connaissance de R,
- Connaissances de base en programmation sont un plus.

## Bibliographie


- James G., Witten D., Hastie T., Tibshiran R., An Introduction to Statistical Learning with Applications in R, Springer, 2015.

— Bottou, Léon, Stochastic learning, Advanced lectures on machine learning, Springer Berlin Heidelberg, 2004. 146-168.
— Package parallel documentation
— Dean, Jeffrey, and Sanjay Ghemawat, MapReduce : simplified data processing on large clusters, Communications of the ACM 51.1 (2008) : 107-113.
— White, Tom. Hadoop : The definitive guide. 4th edition, O'Reilly Media, Inc., 2012.
— Package rmr2 documentation.
— Stochastic Gradient Methods for Large-Scale Machine Learning, L. Bottou, F. E. Curtis, and J. Nocedal, ICML 2016 tutorial.
— Stochastic Optimization for Big Data Analytics : Algorithms and Library, Tianbao Yang, Rong Jin and Shenghuo Zhu, SIAM-SDM 2014 Tutorial.
— Learning With Large Datasets, Andrew Ng - Coursera, online : https://www.coursera.org/learn/machine-learning/lecture/CipHf/learning-with-large-datasets.
— Package parklyr documentation.
— Karau, Holden, et al. Learning spark : lightning-fast big data analysis, O'Reilly Media, Inc., 2015.
— Spark documentation : http://spark.apache.org/docs/latest/index.html
— rstudio/spark documentation : http://spark.rstudio.com/
— Apache Spark Tutorial, Future Cloud Summer School, Paco Nathan, 2015. http://cdn.liber118.com/workshop/fcss_spark.pdf

# INF2204
# Syst. d'information décisionnels et entrepôt de données

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

En informatique décisionnelle, on est amené à traiter de grands ensembles de données, provenant de sources hétérogènes diffuses internes ou externes à l'entreprise. Ces données sont stockées dans des entrepôts, organisées par 'métiers' et décrites suivant des dimensions ou axes d'analyse. Cet enseignement a pour but d'apporter les éléments pour :

- connaître les principaux composants d'un système décisionnel
- savoir concevoir et modéliser un entrepôt de données
- appréhender les différents outils de l'informatique décisionnelle

## Contenu

- Architecture et composants d'un système décisionnel
- Modélisation dimensionnelle des données : faits, dimensions, schémas en étoile et extensions
- Administration des données de l'entrepôt
  - Alimentation de l'entrepôt : outils ETL
  - Qualité des données
  - Métadonnées et référentiel de données
- Organisation et stockage des données dans l'entrepôt
  - Socle, historisation, agrégats, magasins de données (datamarts)
  - Optimisation : gestion des agrégats, parallélisme, fragmentation
  - Structures multidimensionnelles et OLAP

## Prérequis

Connaissance en système d'information et d'un SGBD [INF2105].

## Bibliographie


- Le système d'information décisionnel. Pascal Muckenhirn. Hermès – Lavoisier, 2003
- Building the data warehouse, William H. Inmon, Wiley Editions, 2005
- Le data warehouse, guide de conduite de projet, Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, Eyrolles, 2005
- Oracle Data Warehouse Tuning for 10g, Gavin Powell. Elsevier, 2005
- Business Intelligence avec SQL Server 2005, Bertrand Burquir, Dunod, 2007


Fiche mise à jour le 08/10/2020

# STA2123
# Modèles de durées et Analyse de Survie

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

Le cadre de l'épidémiologie est succinctement abordé d'un point de vue analytique dans des populations exposées à certains types de maladie sur des zones géographiques. La modélisation intervient ensuite plus spécifiquement sur la variable durée de vie. Les modèles permettent à la fois de comparer des comportements avec ou sans facteurs d'exposition et de faire des prédictions en termes de probabilités de survie.

## Contenu

- Eléments d'épidémiologie
- Les concepts de base sur les durées de vie
- Les modèles de base : exponentiel, Weibull, le modèles de valeurs extrêmes, le model Gamma et log-Gamma.
- Modèles de mélange.
- La censure et les modèles statistiques. Types de censure. Censure aléatoire. L'inférence statistique pour les modèles de censure.
- Méthodes non paramétriques
- Comparaisons de groupes de Survie
- Modèles de Survie paramétriques
- Le modèle de Cox
- Adéquation des modèles de Survie
- Généralisation

## Prérequis

Optimisation, Inférence statistique.

## Bibliographie


- Hill C., Com Nougé C., Kramar A., Moreau T. et al., Analyse Statistique des Données de Survie, Médecine-Sciences Flammarion, 1996.
- Klein J. & Moeschberger M., Survival Analysis, Springer, 2003.


Fiche mise à jour le 08/10/2020

# MIS2251
# Projet tutoré

## Modalités pédagogiques

Projet personnel sous la direction d'un enseignant du master. Durée 20 semaines minimum.

## Objectifs

Le projet tuteuré consiste en un travail scientifique personnel à effectuer sous la responsabilité d'un enseignant-tuteur qui a proposé le sujet choisi par l'étudiant. Le sujet et les modalités d'exécution du projet peuvent être variables ; ils sont définis par l'enseignant-tuteur en accord avec l'étudiant et le directeur des études en prenant notamment en compte les souhaits de poursuite d'études de l'étudiant ainsi que son projet professionnel.

## Contenu

Face à un besoin exprimé par l'entreprise, le groupe projet doit, dans un premier temps, proposer une solution et l'organisation à mettre en oeuvre pour la réaliser puis, une fois la proposition validée par l'entreprise, la réaliser. Un soin particulier est apporté pour former les étudiants aux meilleures pratiques du monde de l'entreprise en particulier sur la communication maîtrise d'oeuvre - maîtrise d'ouvrage ainsi que sur le respect des engagements pris. Les projets sont en général l'occasion pour les étudiants de réaliser un projet de sa phase d'analyse à sa réception d'approfondir ou découvrir les méthodes et technologies nécessaires à la réalisation du projet. L'évaluation est faite sur la qualité des livrables et la gestion du projet.

Au début du Semestre 2, les étudiants doivent contacter leur directeur des études ou un enseignant du master afin de choisir leur sujet. Des réunions régulières entre l'étudiant et son tuteur permettent la bonne avancée du projet. A la fin du Semestre 2, les étudiants rendent un rapport scientifique et soutiennent publiquement leur travail.

## Prérequis

Licence de mathématiques et premier semestre du M1.

## Bibliographie

La bibliographie adaptée au projet sera communiquée par le tuteur.

Fiche mise à jour le 08/10/2020

# STA2124
# Optimisation statistique et Business Intelligence

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

— Acquérir une vision d'ensemble des méthodes d'optimisation différentielle utilisées en Statistique et les mettre en oeuvres sur des exemples concrets à l'aide de R, SAS et Python.
— Avoir un premier aperçu des méthodes d'optimisation combinatoire et de leurs applications en Business Intelligence.

## Contenu

— Introduction générale
— Le problème d'optimisation
— Quelques exemples statistiques
— Principe des algorithmes de minimisation
— Définition de la Business Intelligence
— Modélisation en Business Intelligence
— Optimisation en Business Intelligence

## Prérequis

Statistique inférentielle, Méthode du Maximum de Vraisemblance, Programmation R, Programmation SAS.

## Bibliographie


— Optimization Techniques in Statistics. J. Rustagi, Academic Press 1994.
— Introduction to Optimization Methods and Their Application in Statistics. B. S. Everitt, Chapman and Hall 1987.
— Introduction à l'Analyse Matricielle et à l'Optimisation. P. G. Ciarlet, Dunod 1982.
— Fundamentals of Business Intelligence, W. Grossmann, S. Rinder-Ma. Springer 2015.


Fiche mise à jour le 08/10/2020

# SUCG402
# Enseignement complémentaire

## Modalités pédagogiques

Matière de l'UE d'Enseignement Complémentaire / Culture Générale. Cours : 6 h, TD : 10 h en présentiel.

## Objectifs

- Connaître les entreprises qui emploient dans les disciplines des étudiants (mathématique, informatique, statistique)
- Cibler un projet professionnel et mettre au point des stratégies de recherche d'emploi : CV, lettre, entretiens, questionnaires
- Aborder l'actualité et des questions pratiques pour les jeunes diplômés en milieu professionnel

## Contenu

- Présentation des méthodes de recrutement, du marché du travail, de la recherche d'emploi, des ressources
- Ateliers de ré écriture de CV et lettres de motivation
- Exposés étudiants sur des thèmes d'actualité en RH (harcèlement, gestion du stress, donner sa démission...)
- Simulation d'entretiens d'embauche

## Prérequis

Aucun.

## Bibliographie

Webographie de sites d'emploi et de conseils donnée en cours.

Fiche mise à jour le 08/10/2020

# UNITÉS D'ENSEIGNEMENT DU SEMESTRE 3

# STA2326
# Modélisation de données complexes

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

L'objectif de cette UE est de faire découvrir aux étudiants les problèmes posés par l'analyse de données complexes (modélisation statistique, assimilation de données, couplage modèle-données).

## Contenu

1. Données standard et données complexes,
2. Prise en compte de la structure des données
3. Couplage Modèle-Données
4. Estimation à noyau
5. Régression non paramétrique fonctionnelle (design fixe et aléatoire) : estimateur à noyau, régression spline, polynômes locaux
6. Algorithme stochastique
7. Propriétés asymptotiques et comparaison des estimateurs
8. Régression robuste : régression quantile
9. Analyse de données simulées et données réelles

## Prérequis

Statistique mathématique, processus stochastique, Modèle linéaire généralisé, régression linéaire.

## Bibliographie


— Bercu B., Capderou S. and Durrieu G. (2019) A nonparametric statistical procedure for the detection of marine pollution, Journal of Applied Statistics, 46(1), 119-140.
— Bercu B., Capderou S., and **Durrieu G.** Nonparametric recursive estimation of the derivative of the regression function with application to sea shores water quality, Statistical Inference for Stochastic Processes, 22(1), 17-40 (2019).
— Duflo M (1997) Random iterative models, vol 34 of Applications of mathematics (New York), Springer, Berlin
— Fan, J., and I. Gijbels. 1996. Local Polynomial Modelling and Its Applications. Vol. 66. Monographs on Statistics and Applied Probability. London : Chapman & Hall.
— Nadaraya E.A. (1964) On estimating regression, Theory Probab. Appl. 9,141–142.
— Watson G.S. (1964) Smooth regression analysis, Sankhya 26, 359–372.


Fiche mise à jour le 08/10/2020

# STA2328
# Intelligence Artificielle et Deep Learning

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

Ce cours a pour objectif de vous introduire le principe des réseaux de neurones et de l'apprentissage profond. Pour ce faire, la première partie permettra de revoir les concepts de bases et le vocabulaire du Machine Learning - la suite du cours sera alors consacrée à la description des réseaux de neurones afin de les comprendre et de savoir les implémenter (sous Python via Keras/Tensorflow) pour la résolution de problèmes concrets d'analyse de données.

## Contenu

1. Introduction
2. Machine Learning : Fundamental concepts
3. Feedforward neural networks
4. Convolutional neural network (CNN)
5. Recurrent neural network (RNN)

## Prérequis

Bases de l'apprentissage statistique, langage de programmation scientifique (Python).

## Bibliographie


— Goodfellow, I., Bengio, Y. & Courville, A., 2016. Deep Learning, The MIT Press.
— Buduma, N. & Locascio, N., 2017. Fundamentals of Deep Learning, O'Reilly Media.
— Chollet, F., 2017. Deep Learning with Python, Manning Publications.


Fiche mise à jour le 08/10/2020

# STA2325
# Statistique spatiale et SIG

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

— prendre conscience des possibilités de gestion et d'analyse des données géospatiales ou géoréférencées,
— expliquer les principales méthodes et technologies de gestion de ces données,
— utiliser ces méthodes et outils.

## Contenu

1. Introduction aux SIG et leur fonctionnalités.
   — Techniques d'acquisition et d'importation de données, le géo-référencement.
   — Stockage vectoriel et matriciel.
   — Représentations terrestres, projections.
2. Utilisation
   — Requêtes et aide à la décision : fonctions de gestion et d'analyse spatiale.
   — Calage de carte
   — Restitution : analyse des relations spatiales et qualité des données, analyse thématique, méthodes quantitatives. Représentation (analyse en 3D) et sémiologie graphique.
3. Le mode réseau et Application Web.
4. Outils : Arcview, Mapinfo et Geoconcept, logiciels libres.

## Prérequis

Connaissance des systèmes de gestion de bases de données

## Bibliographie

— Robert Laurini et Derek Thompson, Fundamentals of spatial information systems, Academic Press, 1992.

Fiche mise à jour le 08/10/2020

# STA2329
# Challenge Kaggle & Big Data (Hadoop, Spark)

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel. ECTS : 5

## Objectifs

L'objectif de cette UE est d'initier les étudiants aux problématiques liées à l'analyse de grande masse de données (Big data) avec la présentation des bases NoSQL (Cassandra, Neo4j, Redis, MongoDB, HBase), des plateformes BigData (Hadoop, Spark) et du Cloud Computing (Azure, Google, Amazon). La moitié de l'UE est dédiée à une mise en situation concrète d'analyse de données réelles sous la forme d'un challenge de type Kaggle.

## Contenu

1. analyse de données
2. nettoyage des données
3. sélection des features et des modèles
4. les bases NoSQL (Cassandra, Neo4j, Redis, MongoDB, HBase),
5. le BigData (Hadoop, Spark)
6. le Cloud Computing (Azure, Google, Amazon)

## Prérequis

Analyse de données, machine learning, Modèle linéaire généralisé.

## Bibliographie


— Bruce, P., & Bruce, A. (2017). Practical Statistics for Data Scientists : 50 Essential Concepts (pp. 1-562). O'Reilly Media.
— Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow (2nd ed., pp. 1-510). O'Reilly.
— Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning (pp. 1-802). The MIT Press.
— Marr, B. (2015). Big Data : Using smart Big Data, Analytics and metrics to make better decisions and Improve Performance. Wiley.
— Marz, N & Warren, J. (2015) Big Data : Principles and best practices of scalable realtime data systems. Manning Publications.
— McKinney, W. (2012). Python for Data Analysis (1st ed., pp. 1-470). O'Reilly Media.


Fiche mise à jour le 08/10/2020

# STA2321
# Machines à Vecteur Support et méthodes à noyaux

## Modalités pédagogiques

Cours (20h) et TD (22h) en présentiel.

## Objectifs

En statistique la reconnaissance des formes a pour but de détecter et de caractériser les relations entre les données. Les méthodes à noyaux qui sont apparues sous la forme de «machine à vecteur support» (SVM) pour les problèmes de classification, se sont rapidement étendues à d'autres problèmes de la statistique. Il s'agit d'un progrès important dans l'étude des problèmes liés à tous les types de traitement de données, en particulier à des données hautement mutivariées, géoréférencées et souvent longitudinales. L'objectif de ce cours est de proposer des outils de modélisation pour ces données qui permettront de réaliser des simulations et/ou des prévisions.

## Contenu

- Rappel en optimisation convexe sous contraintes. Machine à vecteur support pour la classification : données séparables et non séparables
- L'astuce du noyaux. Machine à vecteur support dans l'espace des traits
- Machine à vecteur support pour la régression
- Analyse des formes par décompositions propres. Décomposition en valeurs singulières (SVD). Méthode à noyaux pour la régression linéaire et la régression PLS. Analyse discriminante de Fisher, ACP avec la méthode à noyaux
- Quelques problèmes de données spatiales
- Données référencées ponctuellement. Données référencées par des régions
- Modèles hiérarchiques pour les données spatiales univariées. Données spatiales multivariées
- Modèles spatiaux pour la survie

## Prérequis

Modèles linéaires généralisé, Analyse discriminante, Survie.

## Bibliographie


- J. Shawe Taylor & N. Cristianini, (2004), Kernel methods for pattern analysis, Cambridge University Press.
- R. Duda, P. Hart & D. Stork, (2001), Pattern classification, Wiley.S. Banerjee, A. Gelfand, B.P. Carlin, Hierarchical Modeling and Analysis for Spatial Data (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).
- L. Rabiner. A tutorial on hidden markov model and selected applications in speech. Proccedings of the IEEE, 77(2) :257– 285, 1989.


Fiche mise à jour le 08/10/2020

# CON2302
# Conférences

## Modalités pédagogiques

Une dizaine de conférences en présentiel.

## Objectifs

Le premier objectif est qu'à travers des présentations faites par des conférenciers - anciens élèves, chercheurs spécialistes des applications des statistiques, acteurs du monde de l'entreprise - l'étudiant découvre comment les statistiques enseignées dans les autres cours du Master interviennent dans des applications concrètes. L'autre aspect important est de faire connaître aux étudiants les différents métiers des statistiques, et les poursuites d'études possibles (thèse et notamment dispositif CIFRE, ingénieur en mathématique, chercheur, etc.) ainsi que de l'aider dans sa recherche de stage, puis d'emploi à l'issu du Master.

## Contenu

Une dizaine de conférences de deux heures. Intervenants prévus chaque année.

Fiche mise à jour le 08/10/2020

# COM2326
# Enseignement complémentaire

## Modalités pédagogiques

Matière de l'UE d'Enseignement Complémentaire / Culture Générale. Cours : 6 h, TD : 10 h en présentiel.

## Objectifs

Anglais

## Contenu

Anglais

## Prérequis

## Bibliographie

Fiche mise à jour le 08/10/2020

# UNITÉS D'ENSEIGNEMENT DU SEMESTRE 4

# STA2324/INF2402
# Stage M2

## Modalités pédagogiques

Stage de 20 semaines à 6 mois, en entreprise ou en laboratoire de recherche.

## Objectifs

Le stage en entreprise est l'occasion de se confronter à la vie professionnelle et de mettre en pratiques les connaissances et savoir-faire théoriques acquis. Le stage en laboratoire est l'opportunité d'approfondir un sujet actuel de recherche.

## Contenu

Le stage se déroule entre mi-janvier et le mois de juin et doit avoir une durée effective minimum de 20 semaines. Chaque stage est suivi par une tuteur enseignant qui effectue si possible une visite en cours de stage. Il est évalué sur plusieurs délivrables, une proposition de projet, due un mois après le début du stage, un pré-rapport à mi-stage et un rapport final, et une soutenance orale publique. La recherche des stages est sous la responsabilité des étudiants ; une commission du Master évalue l'intérêt pédagogique des stages proposés avant leur affectation définitive.

Fiche mise à jour le 08/10/2020