**“Advanced Statistics and Data Analysis”**

**Syllabus CT.02/1**

**Prof. Kyandoghere Kyamakya**

**Prof. Jean Chamberlain Chedjou**

October, 11th, 2018

Version 1

# Name of the course

**Advanced Statistics and Data Analysis**

# ECTS credits

6 Credits, (**45 hours of Theory + 30 hours of Exercises & Lab**), 1 semester

# Objectives

This course familiarizes students with the methods, concepts, techniques and algorithms to analyze systems undergoing random/stochastic dynamics. The analysis is based on both modeling and simulation. Various probability distribution functions are used to model the underlined randomness mathematically. Each distribution corresponds to a specific random-scenario or a specific random-behavior as the degree of randomness might vary from one scenario to another. Specifically, univariate and multivariate distributions are used to model the underlined stochasticity. Therefore, for various scenarios selected in the fields of road transportation, railway transportation, and supply chain networks and logistics we develop a systematical analytical concept leading to the derivation of the corresponding mathematical models expressed in the form on probability distribution functions. The resulting models are further simulated using the scientific computing software MATALAB. Therefore, students are familiarized with the use of MATLAB for the computing of stochastic dynamics.

Further, several methods and techniques are developed and are presented for the analysis of stochastic data. The methods developed are further used to analyze various real-world scenarios selected in three main fields of research namely in: Supply chain networks-and-logistics, Railway transportation, and Road transportation.

Overall, the lecture provides to students a solid background in the modelling and simulation of stochastic phenomena as well as the analysis of stochastic data. The modelling and simulation backgrounds gained by students are further applied in real-world scenarios selected in the aforementioned three main fields of transportation.

# Learning outcomes

The general expectation regarding the knowledge to be provided/acquired is as follows:

* Mastering of the tools/instruments to model stochastic phenomena
* Mastering of the tools for the simulation of stochastic dynamics/behavior
* Familiarizing students with the use of probability distribution functions (PDF) as well as cumulative distribution functions (CDF) in cases of univariate and multivariate distributions.
* Mastering of how to apply the background of modeling and simulation acquired in this lecture to real-world stochastic systems or scenarios selected in three different fields of transportation.
* Mastering the most relevant statistical methods to analyze stochastic data.
* Mastering of the statistical techniques to analyze experimental stochastic data.

# Contents

1. **General introduction**
   1. Importance of statistics and data analysis in road and railway transportation as well as in supply chain networks
   2. Definition of some important keywords and their illustration through concrete examples: *Statistics, Advanced statistics, Data, Data analysis, deterministic, stochastic, distribution functions, mean/average, data forecasting*
   3. Commonly used methods, concepts and algorithms for **data analysis**
   4. Commonly used methods, concepts and algorithms for **data forecasting**
2. **Statistical analysis of stochastic phenomena**
   1. Deterministic systems versus stochastic systems, and some illustrative examples in transportation
   2. Deterministic formalism versus stochastic formalism
   3. Importance and essence of the estimation process in data analysis
   4. Fundamental parameters of a stochastic process and measurements
      1. Mean; Mode/Top ; Median; Mode; Variance/Covariance; Standard deviation;
      2. Percentile; Quantile; Confidence level ; Confidence interval
   5. Importance of the metric Z-score in data analysis
   6. Importance of the metric confidence interval in data analysis
   7. Level of confidence and factors affecting the confidence interval range
   8. Z-score tables: Description, reading techniques, and importance in data analysis
   9. Elements of the confidence interval estimation
      1. The confidence level
      2. The mean/average
      3. The standard deviation
      4. The sample size
      5. The confidence limits (Lower and upper limits)
   10. The central limit theorem (CLT)
   11. Normal distribution versus standard normal distribution
   12. Factors affecting the confidence interval (CI) range: *Data variation; Sample size; Level of confidence; Width of CI*
   13. Confidence interval estimates under some assumptions
   14. Application: ***Analysis of various sets of data and estimation of the confidence interval***
3. **Basics of traffic theory: Fundamentals of queuing and simulation of queuing processes in stochastic scenarios/Events**
   1. Definition of keywords: *traffic; traffic theory; variables; random variables; distribution functions; poisson process; homogeneous poisson processes; non-homogeneous poisson processes; overview of traffic processes.*
   2. Methods used in traffic theory: Pros/Advantages and Cons/drawbacks
   3. Models used in stochastic theory: Pros/Advantages and Cons/drawbacks
   4. Some mathematical probability distribution functions
      1. Exponential
      2. Shifted exponential
      3. Poisson
      4. Erlang
      5. General
   5. Registration equipment
   6. Overview of queuing and applications in transportation
   7. General queuing notation (Ref. Kendall 1951)
   8. Some queuing systems/models
   9. State analysis of queue systems/models
   10. Blocking queuing systems/models
       1. Erlang traffic models: Advantages and limitations
       2. Erlang B; Extended Erlang B; Erlang C
       3. Busy Hour Traffic (BHT) – Blockings – Number of communication lines – Recall factor
       4. Queuing models in VoIP
       5. Main phases of a queuing process
       6. Elements of a queuing system
       7. Sample applications of queuing
   11. General queuing notation (Kendall 1951)
   12. Queuing models
   13. State analysis of queue models/systems
   14. Mathematical modeling of a single-server queuing system
4. **Approximation and ﬁtting**
   1. Norm approximation
      1. Basic norm approximation
      2. Penalty function approximation
      3. Approximation with constraints
   2. Least-norm problems
   3. Regularized approximation
      1. Bi-criterion formulation
      2. Regularization
      3. Reconstruction, smoothing, and de-noising
   4. Robust approximation
      1. Stochastic robust approximation
      2. Worst-case robust approximation
   5. Function ﬁtting and interpolation
      1. Function families
      2. Constraints
      3. Fitting and interpolation problems
      4. Sparse descriptions and basis pursuit
      5. Interpolation with convex functions
   6. Application exercises
5. **Statistical estimation**
   1. Parametric distribution estimation
      1. Maximum likelihood estimation
      2. Maximum a posteriori probability estimation
   2. Nonparametric distribution estimation
   3. Optimal detector design and hypothesis testing
      1. Deterministic and randomized detectors
      2. Detection probability matrix
      3. Optimal detector design
      4. Multicriteria formulation and scalarization
      5. Binary hypothesis testing
      6. Robust detectors
   4. Chebyshev and Chernoﬀ bounds
      1. Chebyshev bounds
      2. Chernoﬀ bounds
      3. Example
   5. Experiment design
   6. Application exercises
   7. Application exercises
6. **Geometric problems**
   1. Projection on a set
      1. Projecting a point on a convex set
      2. Separating a point and a convex set
      3. Projection and separation via indicator and support functions
   2. Distance between sets
      1. Computing the distance between convex sets
      2. Separating convex sets
      3. Distance and separation via indicator and support functions
   3. Euclidean distance and angle problems
      1. Gram matrix and realizability
      2. Problems involving angles only
      3. Euclidean distance problems
   4. Extremal volume ellipsoids
      1. The Löwner-John ellipsoid
      2. Maximum volume inscribed ellipsoid
      3. Aﬃne invariance of extremal volume ellipsoids
   5. Centering
      1. Chebyshev center
      2. Maximum volume ellipsoid center
      3. Analytic center of a set of inequalities
   6. Classiﬁcation
      1. Linear discrimination
      2. Nonlinear discrimination
   7. Placement and location
      1. Linear facility location problems
      2. Placement constraints
      3. Nonlinear facility location problems
      4. Location problems with path constraints
   8. Floor planning
      1. Relative positioning constraints
      2. Floor planning via convex optimization
      3. Floor planning via geometric programming
7. **Selected real-world scenarios as application examples in transportation** 
   1. **Supply Chain Networks** 
      1. Dynamic supply chains with stochastic policies
      2. Dynamic supply chains with stochastic demands
      3. Modelling of a supply chain network driven by stochastic fluctuations
   2. **Railway transportation**
      1. Stochastic analysis of dynamic interaction between train and railway turnout
      2. Simulation of Train Track Interaction with Stochastic Track Properties
      3. Stochastic modeling of track irregularities using experimental measurements
   3. **Road transportation**
      1. Application of Stochastic Modeling and Simulation to Vehicle System Dynamics
      2. Stochastic Modeling and Simulation of Traffic Flow
      3. Stochastic modelling of traffic flow and corresponding models
      4. Traffic flow theory and chaotic behavior

# Teaching method

Lectures, Case studies, Tutorials/exercises, Numerical coding using MATLAB.

* The slides are available for the whole lecture. These slides are must be provided to students (or must be uploaded in the MOODLE system). The full content of each slide is systematically explained by the Lecturer. Additional examples which are not included in slides will be proposed by the Lecturer to allow good understanding of the information provided.
* The slides contain exercises with solutions for the good understanding of the content of each chapter. These solutions are systematically explained (during the lecture) by the Lecturer.
* The Slides contain exercises without solutions to be solved by students during the lecture (this is part of oral exam). The students are fully assisted by the Lecturer in order to obtain correct/exact solutions to the proposed exercises. This will help to check whether the students have understood the chapters or not.
* Several exercises will be proposed by the Lecturer to be solved by students as projects. This will help to test the self-learning potential of students.

# Assessment method

Mid-term and final oral and/or written examination, exercises from case studies.

# Textbooks - Publications - Software

**Textbooks**

* Boyd S., and Vandenberghe L.(2004). **Convex Optimization**, Cambridge University Press, New York Seventh edition, ISBN 978-0-521-83378-3
* Arthur F. and Veinott, Jr (2005). **Supply-Chain Optimization**, Lecture Notes, MS&E 361. Copyright@2005
* **Understanding Advanced Statistics Methods**:   
  https://www.amazon.com/Understanding-Advanced-Statistical-Methods-Chapman/dp/1466512105

**Selected relevant Publications**

* Scheepmaker, G. M., Goverde, R. M. P., Kroon, L. G., 2017. Review of energy efﬁcient train control and timetabling. European Journal of Operational Research vol. 257 (2), 355–376.
* Wang, Y., De Schutter, B., van den Boom, T. J., Ning, B., 2013. Optimal trajectory planning for trains – a pseudospectral method and a mixed integer linear programming approach. Transportation Research Part C: Emerging Technologies vol. 29, 97–114.
* Wang, Y., De Schutter, B., van den Boom, T. J., Ning, B., 2014. Optimal trajectory planning for trains under ﬁxed and moving signaling systems using mixed integer linear programming. Control Engineering Practice vol. 22, 44–56.
* Yang, X., Chen, A., Ning, B., Tang, T., 2016a. A stochastic model for the integrated optimization on metro timetable and speed proﬁle with uncertain train mass. Transportation Research Part B: Methodological vol. 91, 424–445.
* Yang, X., Li, X., Ning, B., Tang, T., 2016b. A survey on energy-efﬁcient train operation for urban rail transit. IEEE Transactions on Intelligent Transportation Systems vol. 17 (1), 2–13.
* Ye,H.,Liu,R.,2016. A multi phase optimal control method for multi-train control and scheduling on railway lines. Transportation Research Part B: Methodological vol. 93, 377–393.
* Ye, H., Liu, R., 2017. Nonlinear programming methods based on closed-form expressions for optimal train control. Transportation Research Part C: Emerging Technologies vol. 82, 102–123.
* Zhou, L., Tong, L. C., Chen, J., Tang, J., Zhou, X., 2017. Joint optimization of high speed train time tables and speed proﬁles: A uniﬁed modeling approach using space time-speed grid networks.Transportation Research PartB: Methodological vol. 97,157– 181.

Software

**\*** Numerical Computing: Algorithmic/Coding and Model-Based Design

* **MATLAB TAH FULL SUITE** for scientific computing, MathWorks (1984), USA.

<https://de.mathworks.com/academia/student_version.html>

* **SPSS** (a software tool for Statistics)