**Course Syllabus for «Programming for Urban Data Analysis»**

Approved by the Academic Council of
_____ Degree Programme[1]
Minutes No. , dated ___ 20____

| | |
|---|---|
| Developer | **Egor A. Kotov,** Research Fellow, Vysokovsky Graduate School of Urbanism, Faculty of Urban and Regional Development, National Research University «Higher School of Economics», ekotov@hse.ru |
| No. of credits | 4 |
| Contact hours | 54 |
| Independent study (hours) | 98 |
| Year of study, degree programme | Master's programme «Urban Development and Spatial Planning», second year |
| Study format | No use of on-line courses |

# 1. Prerequisites, Objectives and Expected Course Results

### 1.1. Prerequisites

The course requires students:

- to be familiar with basic statistics (probability, normal distribution, correlation, univariate and multivariate regression, ANOVA);
- to have basic computer literacy (working with files, folders and archives; understanding absolute and relative paths; understanding CSV and MS Excel files).

The following skills are of advantage, but are not required and can be obtained during the course:

- programming skills (any language) – would greatly reduce the initial slope of the learning curve;
- basic GIS (Urban Planning students that have successfully passed Methods of Spatial Analysis using desktop GIS will have advantage in understanding parts of the course dedicated to working with spatial data);
- working with APIs (Application Programming Interfaces).

### 1.2. Course type

This course is optional for Urban Planning students. It is recommended for those who intend to employ quantitative spatial analysis methods in their term papers and/or masters theses.

### 1.3. Course abstract

Contemporary urban planner and researcher should be aware of the processes that can be observed with new data sources and analysis tools. In the modern urbanised world, enormous amounts of data are generated

[1] *For syllabi from the university-wide pool – head of Department*

daily ranging from citizen complaints and reports to their search queries, daily movements, electricity meter readings, etc. Analysing that data creates new opportunities for studying urban phenomena and enables new scientific approaches in urban planning and management. The extraordinary volume and multidimensionality of urban data require learning new tools and methods for collecting and acquiring such data, shaping it into a specific form appropriate for the analysis, and performing the analysis.

The course introduces the students to the types of data (especially spatial data) relevant to urban research, the advanced tools of working with such data, the full process of data analysis from data collection and exploratory visualisation to inferences, conclusions, presentation of the analysis results. Specific topics include data acquisition, data manipulation and preparation, exploratory analysis, statistical analysis (basic regression and introduction to spatial autocorrelation and regression), data visualisation and reproducible reporting. The students will use R statistical programming language and RStudio IDE (integrated development environment) during the course, but the concepts used in the course and the acquired skills can be applied in Python, Julia or any other programming language with data analysis libraries.

### 1.4. Course Learning Objectives

The objectives of the course are to:

- Familiarise students with different types of urban data sources, file and database types used for storage of such data.
- Discuss the origins and associated limitations of various urban data sources.
- Showcase the practices of explanatory data visualisation in urban planning and research.
- Explain the importance of time and space dimensions of urban data.
- Explain how the data is stored and structured.
- Showcase applications of urban data in real world scenarios (both business and government).
- Develop basic skills of applying statistical analysis to large and small data sets.
- Teach basic principles of exploratory data analysis.
- Show how to communicate urban data analysis results through explanatory data visualisation.

### 1.5. Expected Learning Outcomes

After completing the course students **should be able to**:

- Write readable and error-free data analysis code in R that allows a third party to reproduce and interpret the analysis.
- Acquire spatial urban data from files, remote servers and databases using R packages, API and web-scraping.
- Clean and Transform spatial urban data to prepare it for exploratory and statistical analysis.
- Apply exploratory data analysis (EDA) to reveal time and space variations and patterns in urban data.
- Apply linear and spatial regression models and clustering to interpret space-time variations and patterns of urban processes.

## 2. Course Contents

### 2.1. Course Plan

- Introduction to Smart Cities and Urban Data
- Introduction to Scripted Data Analysis and Reproducible Research
- Data Visualisation and Exploratory Data Analysis
- Urban Data Types and Sources. Getting Access to Data
- Tidy Data. Data Cleaning and Transformation
- Working with APIs and Web Scraping
- Statistical Modelling
- Spatial Data Analysis and Statistics

### 2.2. Contact and self-study work, expected learning outcomes etc.

| **Topic (course section)** | **Total hrs**<br>LC<br>SM<br>onl/sw | **Expected learning outcomes (ELO) to be assessed** | **Assessment formats** |
|---|---|---|---|
| 1. Introduction to Smart Cities and Urban Data | 1<br>0<br>8 | Acquire spatial urban data from files, remote servers and databases using R packages, API and web-scraping | Exam |
| 2. Introduction to Scripted Data Analysis and Reproducible Research | 1<br>4<br>8 | Write readable and error-free data analysis code in R that allows a third party to reproduce and interpret the analysis. | Lab 1, Exam |
| 3. Data Visualisation and Exploratory Data Analysis | 2<br>4<br>10 | Apply exploratory data analysis (EDA) to reveal time and space variations and patterns in urban data | Lab 2, Exam |
| 4. Urban Data Types and Sources. Getting Access to Data | 2<br>4<br>12 | Acquire spatial urban data from files, remote servers and databases using R packages, API and web-scraping | Lab 3, Exam |
| 5. Tidy Data. Data Cleaning and Transformation | 2<br>4<br>12 | Clean and Transform spatial urban data to prepare it for exploratory and statistical analysis. | Lab 4, Exam |
| 6. Working with APIs and Web Scraping | 2<br>8 | Acquire spatial urban data from files, remote servers and | Labs 5-6, Exam |

| Topic (course section) | Total hrs | Expected learning outcomes (ELO) to be assessed | Assessment formats |
|---|---|---|---|
| | LC | | |
| | SM | | |
| | onl/sw | | |
| | 16 | databases using R packages, API and web-scraping | |
| 7. Statistical Modelling | 2 | Apply linear and spatial regression models and clustering to interpret space-time variations and patterns of urban processes. | Labs 7-8, Exam |
| | 8 | | |
| | 16 | | |
| 8. Spatial Data Analysis and Statistics | 2 | Apply linear and spatial regression models and clustering to interpret space-time variations and patterns of urban processes. | Labs 9-10, Exam |
| | 8 | | |
| | 16 | | |
| **Hours for types of classes:** | **14** | | |
| | **40** | | |
| | **98** | | |
| **Total hours** | **152** | | |

Course formats:

- LC – lectures;
- SM - seminars/practical courses/ laboratory work;
- Onl. –online lectures and other Internet courses;
- SW – student independent work.

### 2.3. Detailed Course Plan

#### 1. Introduction to Smart Cities and Urban Data

- Smart city as a concept, as a hype, as a marketing phenomenon, as one of the key causes of emergence of urban data. City as a corporation vs. city as a living organism. Adaptation of the city to new technologies.
- Automated data generation and collection. Urban data ubiquity. The origins of urban data. Urban data sources. Traditional urban data (state urban statistics) vs. new data sources.
- Urban data analysis as part of daily routines of urban dwellers, geo-marketing specialists and tech companies. Outcomes of data ubiquity for urban researchers, planners and managers.
- Required skill sets for urban data analysist.

## 2. Introduction to Scripted Data Analysis and Reproducible Research

- Introduction to scripted data analysis. Point-and-click analysis vs. scripted analysis: head-to-head comparison. Using GUI (graphical user interface) dialog windows vs. calling functions. Importance of reproducible research with motivating examples.
- R language as a statistical command line analysis tool. R language as a programming language. Why R. Comparison of R, Python, Julia, and a few other tools.
- Basics of RStudio IDE (Integrated Development Environment). Working with RStudio projects.
- Reproducible research using R, R Markdown, R Markdown Notebooks, flexdashboard.
- Basic plotting in R. Basic functions and routines applied to classic datasets (mtcars, cars, iris, etc.). Basic data import.

## 3. Data Visualisation and Exploratory Data Analysis

- Storytelling with data. Exploratory vs explanatory analysis. Choosing effective visuals for explanatory analysis. Gestalt principles of visual perception. Spotting bad graphs and maps.
- Exploratory data analysis (EDA) process and tools. Plots vs summary statistics.
- Rorschach protocol. Line-up protocol.
- R tools for Exploratory data analysis. Advanced plotting using ggplot2 and associated tools.
- Interactive plots in R, the simple way.
- Plot design layer by layer. Plot customisation according to Gestalt principles of visual perception. Plot optimisation for colour blind accessibility.

## 4. Urban Data Types and Sources. Getting Access to Data

- Data sources. Open data. Code books. Means of accessing the data. Working with multiple data sources. Data storage file formats. Databases. Getting data from databases. Intro to getting data from web sources using APIs.
- Basic types of data, operators, commands, functions. Approaches to working with data using R. Basic data structures. Objects.
- R object types: vectors, matrices, “data.frames”, “tibbles” and “data.tables”. Lists. Differences between object types and use cases.
- Exporting data to various formats. Choice of storage file format depending on storage goals.
- Basic data manipulation using “data.table” and “dplyr”.

## 5. Tidy Data. Data Cleaning and Transformation

- Wide vs long data. Data reshaping and manipulation. Shaping data in analysis-appropriate form.
- Tidy data concept. Data cleaning. String and date manipulation.
- Regular expressions and their applications for data cleaning. Common pitfalls of regular expressions.
- Feature creation. Data type conversion.
- Building algorithms for data processing.
- Creating functions custom functions, conditional statements, loops for data processing and visualisation.

## 6. Working with APIs and Web Scraping

- Advanced work with APIs. Reading API documentation.
- Constructing API requests. Processing API responses. Data manipulation for converting API responses into analysis-appropriate form.
- Building algorithms for automated data retrieval using APIs.
- Web scraping and related copyright and ethical issues.
- Simple web scraping techniques. Reshaping of scraped data into analysis-appropriate form.

### 7. Statistical Modelling

- Correlation. Simple linear regression. Model fit and interpretation.
- Multiple regression. Simple feature selection. Parallel slopes models.
- Simple cluster analysis techniques.
- A unified framework for application of statistical models in R. Visualization of model results and performance.

### 8. Spatial Data Analysis and Statistics

- Basics of working with spatial data in R. Spatial data storage formats and object types. Importing spatial data from various sources.
- Visualising spatial data in R. Static plotting of spatial data. Interactive maps.
- Merging and joining spatial data. Spatial data analysis. Geometric operations.
- Introduction to spatial statistics. Spatial autocorrelation. Spatial segregation. Spatial generalised linear models.

## 3. Assessments

### 3.1. Honour Code

HSE rules on academic misconduct apply to all labs and the exam in this course. Instances of academic misconduct will result in 0 grade for all involved parties.

**During the labs and the exam students are free to consult:**

1) Course lecture slides;
2) Course lab slides;
3) Personal notes form lectures and labs;
4) Personal completed labs, as well as graded labs with tutor's or other students' comments;
5) Solutions for the previous labs provided by the tutor;
6) The Internet: R documentation, Google, Stack Overflow, etc.

Any considerable amount of code copied from Stack Overflow or blog posts or GitHub, etc. MUST be referenced in the code comments or above/below the code chunk in the markdown document, otherwise, it is regarded as an act of academic misconduct.

**Students are NOT allowed to consult each other during the labs or the exam.** All labs and the exam are individual tasks, unless otherwise explicitly stated by the tutor. Collaborative work is prohibited and considered to be an act of academic misconduct. Experience shows that collaborative work on the labs results in mindless replication of mistakes. Collaborative work not only results in reduced grades but prevents the students from thinking on their own (because of the ease of copying the code) and developing the required skills and reaching the expected learning outcomes.

Discussion of the labs is only allowed (and encouraged) between students who have submitted their labs for grading.

### 3.2. Grading and Guidelines for Knowledge Assessment

In general, any lab is to be due in 8 to 15 days after the lab work is handed out to the students, depending on the lab difficulty. Some exceptions may apply due timetable in a given year. Specific deadline dates are announced on the day of lab.

After the deadline of every lab a solution is provided. No completed lab submissions are accepted after that date. Therefore, any student who fails to submit a lab before the deadline gets no grade for it. All labs are interconnected and building on the knowledge of one another, so completing labs on an ad hoc basis and in no particular order is not a good option. For example, beginning lab 3 without completing labs 1 and 2 prior to that would be a challenge, since in lab 3 a student would be expected to know the concepts and tools of the first two labs.

Blocking assessment elements are not present.

#### 3.2.1. Grading System

Labs may be graded through a peer-review process. In this case labs are anonymised by the tutor and distributed to the students for grading. Every enrolled student needs to evaluate a number of labs, otherwise he or she will not be able to get a grade for his or her own lab submission.

Every lab and the exam are evaluated on a 0 to 10 scale adopted by the HSE. The final grade for the course consists of accumulated and exam grades. Rounded is performed with round half up approach, e.g. 8.49 or below rounds down to 8, 8.50 and above rounds up to 9. Detailed lab weights table is presented below.

| **Assessment** | Weight in Final Grade |
|---|---|
| Lab 1 | 0.025 |
| Lab 2 | 0.05 |
| Lab 3 | 0.05 |
| Lab 4 | 0.05 |
| Lab 5 | 0.05 |
| Lab 6 | 0.075 |
| Lab 7 | 0.1 |
| Lab 8 | 0.1 |
| Lab 9 | 0.1 |
| Lab 10 | 0.1 |
| **Exam** | 0.3 |
| **Final Grade (total)** | **1** |

**To express the same with formulae:**

$$Final\ Grade = Round\_Half\_Up(\sum Lab_i * Weight_i + Exam * 0.3\ )$$

Final Grade is not rounded to 4 if the pre-rounded figure is less than 4, e.g. 3.99 is not rounded to 4.

#### 3.2.2. Examination type

The course concludes with an exam in form of a data analysis task based on the concepts and tools covered in the course.

### 3.3. Methods of Instruction

The class generally meets weekly or fortnightly. The class usually starts with a lecture followed by a lab introduction and then a lab work itself. Previous lab works may be discussed and showcased.

Key methods of instruction for the course are:

- lectures;
- demonstration;
- controlled and uncontrolled practical exercises.

# 4. Resources

## 4.1. Core readings

| № | Name |
|---|---|
| 1. | Townsend, A. M. (2014) *Smart cities: big data, civic hackers, and the quest for a new utopia*. New York: W.W. Norton & Company.<br>• Introduction: Urbanization and Ubiquity<br>• Chapter 1: Urbanization and Ubiquity<br>• Chapter 4: The Open-Source Metropolis<br>• Chapter 5: Tinkering Toward Utopia<br>• Chapter 7: Reinventing City Hall<br>• Chapter 8: A Planet of Civic Laboratories |
| 2. | Peng, R.D., 2015. R Programming for Data Science, Available at: https://leanpub.com/rprogramming.<br>• History and Overview of R<br>• Getting Started with R<br>• R Nuts and Bolts<br>• Getting Data In and Out of R<br>• Using the readr Package<br>• Using Textual and Binary Formats for Storing Data<br>• Dates and Times<br>• Control Structures<br>• Functions |
| 3. | Peng, R. D. (2015) Report Writing for Data Science in R. Leanpub. Available at: https://leanpub.com/reportwriting. |
| 4. | Knaflic, C. N. 2015. Storytelling with data: a data visualization guide for business professionals. New Jersey: Wiley, Available at: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4187267 |
| 5. | Burchell, J. and Sepúlveda, M. V. (2017) The Hitchhiker's Guide to Ggplot2. Leanpub. Available at: https://leanpub.com/hitchhikers_ggplot2. |
| 6. | Peng, R.D., 2015. Exploratory Data Analysis with R, Available at: https://leanpub.com/exdata.<br>• The ggplot2 Plotting System: Part 1<br>• The ggplot2 Plotting System: Part 2<br>• Plotting and Color in R |

| № | Name |
|---|---|
| 7. | Wickham, H., 2009. ggplot2, New York, NY: Springer New York. Available at: http://link.springer.com/10.1007/978-0-387-98141-3.<br>• Chapter 1: Introduction<br>• Chapter 4: Build a plot layer by layer (4.1-4.5.2, 4.6-4.9)<br>• Chapter 5: Toolbox (5.1-5.3) |
| 8. | Wickham, H. and others (2014) 'Tidy data', Journal of Statistical Software, 59(10), pp. 1–23, Available at: http://courses.had.co.nz/12-rice-bdsi/slides/07-tidy-data.pdf |
| 9. | Tidy data presentation by Hadley Wickham: http://stat405.had.co.nz/lectures/18-tidy-data.pdf |
| 10. | Munzert, S., Rubba, C. & Nyhuis, D., 2015. Automated Data Collection With R: A Practical Guide to Web Scraping and Text Mining, Chichester: John Wiley & Sons Ltd.<br>• Chapter 8: Regular expressions and essential string functions<br>• Chapter 9: Scraping the Web (9.1.1-9.1.3, 9.2-9.4)<br>• Chapter 14: Predicting the 2014 Academy Awards using Twitter |
| 11. | Schmelzer, C.H., Martin Arnold, Alexander Gerber and Martin, 2018. Introduction to Econometrics with R, Available at: https://www.econometrics-with-r.org/ |
| 12. | Pace, L. and Hlynka, M. (2012) Beginning R an introduction to statistical programming. New York: Apress. Available at: http://www.books24x7.com/marc.asp?bookid=50030 (Accessed: 11 February 2018). |
| 13. | Speegle, D., 2018. Foundations of Statistics with R, Available at: https://bookdown.org/speegled/foundations-of-statistics/ |
| 14. | Lovelace, R., Nowosad, J., Muenchow, J., 2018. Geocomputation with R, Available at: https://geocompr.robinlovelace.net |

### 4.2. Additional readings

| № | Name |
|---|---|
| 1. | Dietmar, O. & Ratti, C., 2014. *Decoding the City: Urbanism in the Age of Big Data* 1st ed., Basel: Birkhauser Verlag AG.<br>• Seeing the City through Data / Seeing Data through the City (by Kael Greco)<br>• The Kind of Problem a City Is: New Perspectives on the Nature of Cities from Complex Systems Theory (by Luís M. A. Bettencourt) |
| 2. | Batty, M., 2007. Complexity in city systems: Understanding, evolution, and design. A planner's encounter with complexity, Available at: http://discovery.ucl.ac.uk/3473/1/3473.pdf |
| 3. | Peng, R. (2015) 'The reproducibility crisis in science: A statistical counterattack', Significance, 12(3), pp. 30–32, Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2015.00827.x/full |
| 4. | Yau, N. (2013) Data points: visualization that means something. Indianapolis, IN: John Wiley & Sons, Inc. |
| 5. | RStudio cheat sheet by RStudio http://www.rstudio.com/wp-content/uploads/2016/01/rstudio-IDE-cheatsheet.pdf |

| № | Name |
|---|---|
| 6. | Online markdown editor 1: http://markdown.pioul.fr/ |
| 7. | Online markdown editor 2: http://dillinger.io/ |
| 8. | Online markdown tutorial 1: http://markdowntutorial.com/ |
| 9. | Online markdown tutorial 2: http://www.markdown-tutorial.com/ |
| 10. | Online markdown tutorial 3: http://commonmark.org/help/ |
| 11. | ggplot2 official online reference: http://docs.ggplot2.org/current/ |
| 12. | ggplot2 graph examples at Plot.ly: https://plot.ly/ggplot2/ |
| 13. | ggplot2 cheat sheet by RStudio: http://www.rstudio.com/wp-content/uploads/2015/12/ggplot2-cheatsheet-2.0.pdf |
| 14. | Illustrated ggplot2 tutorial by Basel Institute for Clinical Epidemiology and Biostatistics: http://www.ceb-institute.org/bbs/wp-content/uploads/2011/09/handout_ggplot2.pdf |
| 15. | Plotly online reference: https://plot.ly/r/ |
| 16. | Plotly online detailed reference: https://plot.ly/r/reference/ |
| 17. | readr reference: https://cran.r-project.org/web/packages/readr/readr.pdf |
| 18. | readxl reference: https://cran.r-project.org/web/packages/readxl/readxl.pdf |
| 19. | openxlsx reference: https://cran.r-project.org/web/packages/openxlsx/openxlsx.pdf |
| 20. | R data.table cheat sheet by DataCamp: https://s3.amazonaws.com/assets.datacamp.com/img/blog/data+table+cheat+sheet.pdf |
| 21. | Discussion on variable assignment using '=' and '<-': http://stackoverflow.com/questions/1741820/assignment-operators-in-r-and |
| 22. | R data.table cheat sheet by DataCamp: https://s3.amazonaws.com/assets.datacamp.com/img/blog/data+table+cheat+sheet.pdf |
| 23. | RegEx online playground 1: http://www.regexr.com/ |
| 24. | RegEx online playground 2: https://regex101.com/ |
| 25. | R data.table cheat sheet by DataCamp: https://s3.amazonaws.com/assets.datacamp.com/img/blog/data+table+cheat+sheet.pdf |
| 26. | rvest tutorial: https://stat4701.github.io/edav/2015/04/02/rvest_tutorial/ |

### 4.3. Special Equipment and Software Support

Classrooms must be equipped with projectors for showing lecture slides and demonstrating lab assignments.

The classrooms must be equipped with the following list of free software:

- R 3.6.x+ ( https://cran.r-project.org/ )
- RStudio 1.2.x+ ( https://www.rstudio.com/products/rstudio/ )

Students who choose to follow the class using private computers/notebooks are responsible for installation of the required software.

## 5. Organization of Studies for Persons with Limited Mobility and Disabilities

If necessary, learners with limited mobility or a disability (as per his/her application), as well as per his/her individual rehabilitation programme, may be offered the following options for receiving learning information with due consideration of his/her individual psycho-physical needs (e.g., via eLearning studies or distance technologies):

- *for persons with impaired vision*: enhanced fonts in hard copy documents; e-documents; audio files (transfer of study materials to an audio-format); hard copy documents with the use of Braille; individual consultation with a facilitated communicator; individual assignments and mentoring;
- *for persons with hearing impairments*: in hard copy; e-documents; video materials with subtitles; individual consultation with a facilitated communicator; individual assignments and mentoring;
- *for persons with a muscular-skeleton disorder*: in hard copy; e-documents; audio-files, individual assignments and mentoring.