**MUSA 5000/CPLN 6710 – Statistical and Data Mining Methods for Urban Data Analysis ∙ Fall 2023**

**Instructor: Eugene Brusilovskiy (*he/him/his*) ∙** [**eugeneby@design.upenn.edu**](mailto:eugeneby@design.upenn.edu) **∙ Cell: 267-939-0230**

**Time: MW 5:15-6:45 ∙ Location: TBD**

***Course Overview and Objectives***

The goal of this course is to familiarize students with a number of statistical and data mining techniques commonly used for analyzing different types of urban data. The course will have two key parts: 1) regression analysis for urban data, and 2) identifying patterns in urban data. Even though the course will cover the theory and mathematics behind the majority of the statistical methods, it will focus mostly on their applications. Examples of how planners and other city officials could apply these methods will be presented. For instance, we will examine urban poverty and its predictors, racial residential segregation in urban areas, and socioeconomic disparities in access to various community resources. Students will learn when each statistical test should be used and the assumptions behind each test. In addition, while the R software will be used for most of the course, students will also learn to use GeoDa, an open source mapping and spatial statistics package. Finally, students will learn how to describe and interpret the output of the analyses that they run.

***Class Notes***

Each lecture will come with notes (generally in MS Power Point, MS Word or R Markdown format) that will contain examples of problems very much like the ones that students will be asked to do in their homework assignments.

***Suggested Textbooks/Bibliography***

This course is not based on any single textbook. Instead, it draws topics and material from many different books and online sources. Students may find the following **optional** books helpful:

1. *Devore: Probability and Statistics, 7th Edition.* **Earlier editions may be used as well.**
2. *James, Witten, Hastie & Tibshirani: Statistical Learning with R.* Available for free at <http://faculty.marshall.usc.edu/gareth-james/ISL/>.
3. *O’Sullivan & Unwin: Geographic Information Analysis, 2nd Edition.* Available for free from the Penn library (in electronic format). I will put the book in the course folder.
4. *Anselin: Exploring Spatial Data with GeoDa™: A Workbook.* Available for free at <http://www.csiss.org/clearinghouse/GeoDa/geodaworkbook.pdf>.
5. *Anselin: Spatial Regression Analysis in R: A Workbook*. Available for free at

<https://dces.wisc.edu/wp-content/uploads/sites/128/2013/08/W14_Anselin2007.pdf>.

1. *Williams: Hands-On Data Science with R.* Available for free at <http://onepager.togaware.com/>.

# Course Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** |  | **Topic** | **Sample Applications** | **Optional Supplemental Readings** |
| ***Introductions*** |  |  |  |  |
| 8/30/2023 |  | *Lecture 1.* Introductions. Course Overview. |  |  |
| **9/4/2023** |  | ***No Class - Labor Day*** |  |  |
| 9/6/2023 |  | *Lecture 2.* Introduction to R |  | - Torfs & Brauer: A Short Introduction to R. March 2014. |
| ***Part 1. Regression Analysis Methods for Urban Data*** | | |  |  |
| 9/11/2023 |  | *Lecture 3.* Introduction to Pearson and Spearman Correlation and Other Measures of Association | Is there a relationship between poverty and race?  Are COVID-19 morbidity/mortality associated with poverty, education, and minority status? | - Devore, Ch. 12: Simple Linear Regression and Correlation |
| 9/13/2023 | *Lectures 4-6.* Simple and Multiple regression. Parameter Estimation. Significance Testing of Parameters. Regression Assumptions. Regression Diagnostics. Regression with Categorical Predictors. | - Devore, Ch. 12: Simple Linear Regression and Correlation  - Devore, Ch. 13: Nonlinear and Multiple Regression  - James, Witten, Hastie, Tibshirani, Ch. 3: Linear Regression |
| 9/18/2023 |
| 9/20/2023 |
| 9/25/2023 | ***No Class – Yom Kippur*** |
| 9/27/2023 | *Lectures 7-8.* Regression with Log Transformations. Polynomial Regression. Stepwise Regression. Cross-validation. K-fold validation. |
| 10/2/2023 |  |
| 10/4/2023 | *Lecture 9.* Ridge and Lasso Regression. | - James, Witten, Hastie, Tibshirani, Ch. 6: Linear Model Selection and Regularization |
| 10/9/2023 | *Lecture 10.* Global and Local Indices of Spatial Autocorrelation. Introduction to GeoDa. | What does racial residential segregation in Philadelphia look like? | - Anselin: GeoDa Workbook |
| 10/11/2023 | ***No Class – Fall Break*** |  |  |
| 10/16/2023 | *Lectures 11-12.* Spatial Lag and Spatial Error Regression Assumptions, Interpretation of Output, Diagnostics. Comparisons with OLS. | What is the relationship between house price, poverty, crime, and space? | - Anselin: Spatial Regression Analysis in R: A Workbook |
| 10/18/2023 |
| 10/23/2023 | *Lectures 13-14.* Geographically Weighted Regression: Spatial non-stationarity and concept of varying Beta coefficients. Comparing OLS, Spatial Lag, Spatial Error and GWR | - O'Sullivan & Unwin, Ch. 8: Local Statistics |
| 10/25/2023 |  |
| 10/30/2023 | *Lectures 15-16.* Introduction to Logistic Regression. Interpretation, Diagnostics, etc. | Examining the predictors of healthcare access | - James, Witten, Hastie & Tibshirani, Ch. 4.3: Logistic Regression |
| 11/1/2023 |
| 11/6/2023 |  | *Lecture 17.* Multinomial Logistic Regression |  |
| 11/8/2023 |  | Lecture 18. Designing Research Studies and Analyzing Data. |  |  |
| 11/13/2023 |  | No Class |  |  |
| ***Part 2. Finding Patterns in Urban Data*** | | |  |  |
| 11/15/2023 |  | *Lectures 19-20.* Point Pattern Analysis. Complete Spatial Randomness. Clustering and Uniformity. Nearest Neighbor Analysis.. | Are healthy food markets clustered in PA? Are minorities at a greater disadvantage when it comes to healthy food access? | - Buyong, Ch. 3: Point Data Analysis; - O'Sullivan & Unwin, Ch. 5: Point Pattern Analysis; - O'Sullivan & Unwin, Ch. 6: Practical Point Pattern Analysis |
| 11/20/2023 |
| 11/22/2023 | ***No Class – Thanksgiving*** |  |  |
| 11/27/2023 | *Lecture 21. Point Pattern Analysis (Continued).* K Functions. Adjusting for Edge Effects. Adjusting for Population Density |  |  |
| 11/29/2023 | *Lecture 22.* K-Means Clustering. Introduction to Hierarchical Clustering and DBSCAN. | TBD |  |
| 12/4/2023 | *Lecture 23.* Big Data/Live Data. Using ChatGPT in R. Analyzing Text with R. An Introduction to Text Mining. | What does the geographic distribution of verbal violence look like? |  |
| 12/6/2023 | *Lecture 25.* A Guest Lecture on Big Data |  |  |
| 12/11/2023 | *Lecture 26.* Review and Class Activity |  |  |

***Video Recording of Lectures***

Whenever possible, some class sessions, including your participation, will being recorded and the recordings will be made available to the class, for the duration of the semester. These recordings are made available solely for your personal, educational use and may not be shared, copied or redistributed without the permission of Penn and the instructor. Note also that you are NOT allowed to record class sessions yourselves.

***Software***

The software required for the course will be installed on the computers in the Meyerson Hall computer labs.

1. MS Excel
2. R (latest version can be downloaded from <https://www.r-project.org/>)
3. RStudio (latest version can be downloaded from <https://rstudio.com/>)
4. ArcGIS (Spatial Statistics Toolbox; Spatial Analyst Extension)
5. GeoDa v. 1.14 (can be downloaded from <http://geodacenter.github.io/index.html>)

***Assignments and Grading***

Six homework assignments, which are collectively worth 90% of the final grade, will be issued. These assignments are listed below. Students are expected to complete 5 of them. **Everyone must complete the first four assignments, and then choose either assignment 5 or 6.**

Each assignment will come with detailed instructions on what students are expected to hand in.

The assignments, due dates and percent of grade are listed below,

1. OLS Regression 20% of your grade Due 10/19/2023

2. Spatial Lag, Spatial Error, and Geographically Weighted Regression 25% of your grade Due 11/9/2023

3. Logistic and Multinomial Logistic Regression 20% of your grade Due 11/21/2023

4. Point Pattern Analysis 15% of your grade Due 12/5/2023

5. K-Means Clustering 10% of your grade Due 12/15/2023

6. Analyzing Twitter Data 10% of your grade Due 12/15/2023

Assignments should be e-mailed to me at [eugeneby@design.upenn.edu](mailto:eugeneby@design.upenn.edu) by 11 PM on the due date. If you cannot complete an assignment on time, you should contact me in advance so that an extension may be granted.

For each of the assignments, students should work **in teams of three** and hand in a single submission for the team. All three students working on a team will receive the same grade. Individual submissions are not permitted except in special circumstances. Groups of four or more students are also not allowed.

Students who receive a grade that’s lower than 80 on an assignment (lower than B-) who wish to revise and resubmit it may do so within a week of receiving the assignment back. The final grade for the assignment will be the average between the grade for the initial and final submissions, but will not exceed 85.

The remaining 10% of the final grade will be based on class participation, as students are strongly encouraged to attend the real-time lectures and to ask questions.

Approximately 50% of the students will receive grades in the ‘A’ range (A+, A, A-), 40% will receive grades in the ‘B’ range (B+, B, B-), and 10% will receive grades below a B-.

The following scale will be used for issuing final grades:

A+     100+ B+     87-89 C+     77-79

A       93-99 B       83-86 C       73-76

A-      90-92 B-      80-82 C-      70-72 F       0-69

In order to receive an A+ grade for the semester, a student needs to receive scores of 100 or higher on *all* assignments on the first attempt.

***Academic Integrity and Use of ChatGPT and Other AI Tools***

It is expected that all students adhere to the University’s Code of Academic Integrity, which may be found here: <https://provost.upenn.edu/policies/pennbook/2013/02/13/code-of-academic-integrity>. Please read the policies and *ask questions* as needed.

The use of AI tools including, but not limited to, ChatGPT is discussed below:

Students are allowed – and even encouraged – to use the AI tools for the following purposes:

* *Clarifying and debugging code for assignments.* Because this is not primarily a programming class, and R code will be provided to students (usually in the form of R Markdowns), students are free to use ChatGPT/AI for any additional help they may need with understanding or debugging the code.
* *Correcting spelling and grammar in a report that they write.* ChatGPT/AI is excellent at editing submitted text for spelling and grammar without making substantive changes. However, students should keep in mind that statistics has a very specific language (e.g., failing to reject a hypothesis rather than accepting a hypothesis) that ChatGPT/AI may edit incorrectly. If ChatGPT/AI is used for these purposes, the original and edited versions of the assignment will need to be submitted.

Students are NOT allowed to use AI tools for the following purposes:

* *Writing any part of the assignments, unless explicitly told otherwise.*
* *Any other purpose not explicitly allowed above.*

If you have a question about whether you’re allowed to use ChatGPT/AI for a specific task, just ask, and I will provide clarification!

***Office Hours***

By appointment. I am ALWAYS happy to meet to answer your questions, so please don’t hesitate to reach out. I’ll make sure to meet at a time that works for you and your time zone.

### Teaching Assistant

Ann Zhang

[annzhang@design.upenn.edu](mailto:annzhang@design.upenn.edu)