Course outline for Geostatistics and Data Analysis (Geostats) 2025 • EPSC 552 (3 credits)

This document is in addition to information provided on the course website (https://eps.mcgill.ca/~hinsberg/Stats/Welcome.html). Where there is a discrepancy in the information provided, the information on the website is to be regarded as correct.

Course prerequisites:

This course does not have formal prerequisites. However, you are expected to have a thorough knowledge of spreadsheet programs and their (statistical) functions, or an equivalent proficiency in a programming environment such as Python or R. This ensures that you have the necessary proficiency in working with data and writing functions to process these data.

Course schedule:

Lectures: Tuesday, 10:00 – 12:00 in FDA 348

Labs: Tuesday from 14:30 to 17:30 in FDA 348

Instructor:

Vincent van Hinsberg Email: Vincent.vanHinsberg@mcgill.ca.

Office hours: Tuesday from 12:00 to 14:30 in office FDA 341. Others days or times by email appointment.

There is no TA assigned to this course this year.

Course overview:

On the use of statistical techniques in the geosciences • Statistical techniques provide powerful tools for analyzing and interpreting data, and in this course, you will become familiar with the most commonly used techniques to analyze data in the geosciences. Starting with basic statistical parameters we will gradually move to more complex multivariate techniques, including cluster analysis, factor analysis and (multiple) regression. Geoscience data almost invariably have a spatial component to them, and we will explore the dedicated statistical methods developed for these, including bubble plots, semi-variance, and kriging. The course will mainly focus on the fields of data analysis and statistical testing and modeling. Some aspects of probability analysis will be addressed as well, mainly in relation to confidence intervals and the concept of "statistical proof".

The course consists of lectures that focus on theory, and practical sessions where the tools introduced in the lectures will be applied to geo-datasets.

Required Course Materials:

The book that we will use for the course is "*An introduction to geological data analysis*" by Swan and Sandilands (Blackwell publishing, ISBN 0632032243). The main advantage of this book over standard statistics textbooks is that it is tuned specifically to techniques relevant to the geosciences. There is an earlier edition of this book. If you decide to get this instead, make sure to download the list of errata from the publisher's website – there are quite a few.

The following sections of the book will be covered in the course and should be studied;

chapter 1: completely

chapter 2: 2.1, 2.2, 2.4 (except 2.4.5.3/4), 2.5, 2.6 up to 2.6.2.3

chapter 3: up to 3.3.2

chapter 4: completely

chapter 7: 7.1 & 7.2 (general concepts only), 7.3, 7.4.3

chapter 8: 8.1.1, 8.1.2, 8.3, 8.4, 8.5, 8.6 (general concepts only)

The lectures provide a comprehensive overview of each of the course topics and you can pass the midterm and final exams without having the book. However, I still highly recommend purchasing a copy as it provides an excellent reference on (geo)statistics. The book is also readily available second-hand from various online retailers (e.g., AbeBooks, Amazon, etc.).

There are also many excellent online resources on statistics and data analysis, including a full reference work by the creators of the NCSS software (https://www.ncss.com/software/ncss/ncss-documentation/).

The labs consist of applying statistical tools to geo-datasets and require a computer and dedicated software. Many statistical parameters and plots can be prepared using a spreadsheet program, but

for more advanced statistical analyses we will use the PAST statistics package (http://folk.uio.no/ohammer/past/). This package is freeware and available for Mac and Windows operating systems. If you do not have access to a Windows or Mac computer, you can make use of the computers in the lounge. It is assumed that you are intimately familiar with handling and analyzing data in spreadsheet programs such as Excel, Calc or QuattroPro (any will work), and setting up formulae and functions in these programs. If you need to refresh this, most spreadsheet programs offer help files, both offline and online (for example: Excel – https://support.office.com/en-gb/excel or Calc – www.linuxtopia.org/online_books/office_guides/openoffice_calc_user_guide/index.html).

Learning outcomes:

The aims of this course are to obtain the theoretical understanding and the practical skills to obtain meaningful information from your data and determine the confidence of interpretations and conclusions. Given the central position of data in our, and associated fields, this is a critical skill.

Specifically, upon successful completion of the course, you will be able to;

- Describe your data using standardized statistical parameters; understand which ones are appropriate for your type of data and why this is important; know how to combine datasets
- Understand data uncertainty; design a data collection protocol that ensures and monitors data quality; assess the quality of your data as well as that reported by others; report data quality using standardized and appropriate parameters; derive confidence estimates for your data, and the interpretations and conclusions derived from them
- Apply any statistical test and interpret its outcome; select the most appropriate test(s) for your data; report outcomes in terms of probabilities and confidence levels
- Apply a wide variety of uni-, bi- and multi-variate statistical methods and know how to select which ones are appropriate for your data; be aware that methods have assumptions and know how to test these assumptions; determine whether results are meaningful
- Analyse specific geoscience data types including directional data (flow, alignment, layer orientation), temporal data (stratigraphy and time-series), and spatial data
- Have a solid foundation for exploring other statistical methods and know the nomenclature and vocabulary to understand these methods and apply them correctly
- Have obtained a thorough understanding of the PAST statistical freeware

The labs will moreover reinforce your **teamwork** skills, including distributing tasks, time manage–ment and discussion of findings, and **scientific report writing**, including adhering to the structure of a scientific report, correct reporting of results, and argumentation of interpretations.

Instructional methods:

The course consists of 2 hours of lectures and 3 hours of labs a week. Lectures will focus on the background and theoretical aspects of geostatistical methods and analysis of geo-data, whereas this

knowledge will be applied in the labs. Indeed, the best approach to learning (geo)statistical tools, interpreting their results, and identifying how and where these tools can aid in understanding data, is in working with real-world geo-data. We will do in this in two ways in the lab component of this course. In order to apply the tools and methods discussed in the lectures, and become familiar with calculating statistical properties in spreadsheet programs and the PAST statistics software, the first set of labs are exercise-based and involve preparing data for analysis, statistical analysis of these data, and interpretation of statistical output. In the second set of labs, generally starting from week 5 of the course, you will statistically interpret a large litho-geochemical dataset of soils collected in BC. The dataset contains geological information, element concentrations and field observations, which will have to be explored in combination. You will explore this dataset as a group and apply the full diversity of statistical methods, tools and approaches discussed in the course. The results of this group project are to be presented in a report that will be graded.

Course Topics:

1. *Univariate data and descriptors*: what are data; types of data; populations and samples; mean - median – mode; histograms; normality; outliers; modality; box-and-whiskers plots; stem and leaf diagrams; violin plots; robust descriptors; Z-scores; data levelling; closure
2. *Measures of uncertainty*: sources of uncertainty; range; standard deviation and standard error; variance; inter-quartile range; MAD; error propagation; Bayesian methods; data QA-QC; duplicates
3. *Missing values*: common problem in geosciences and too often ignored - real missing values vs. detection limits; how to deal with missing values; missing value estimation
4. *Statistical testing*: hypotheses; confidence levels; sample size dependence and t-distribution; value and rank testing; normal and robust tests; Z-, t-, Chi-squared, Kolmogorov-Smirnov, Mann-Whitney tests; ANOVA
5. *Regression and correlation*: bivariate statistics; scatter diagrams; Pearson & Spearman correlation coefficients; significance of correlation; regression analysis; robust regression; assumption testing; goodness-of-fit; curve fitting; (non-)linear models; predictive power
6. *Time series data:* time as special variable; random versus periodic data; Markov-chain analysis; auto-correlation; cross-correlation; Fourier transforms and periodograms
7. *Multivariate techniques*: sum of squares methodology; discriminant function analysis; principle component & factor analysis; partial least squares analysis; hierarchical and partitioning cluster analysis; fuzzy methods; machine learning
8. *Spatial data analysis:* spatial distribution of data; 3D visualization (isolines, worms, bubble plots, trend surfaces); spatial interpolation; nearest neighbour; radius methods; semi-variance; kriging
9. *The statistics of improbable events* (earthquakes, volcanic eruptions, ore deposits); distribution tails; forecasting; decision trees

Means of Assessment:

Title	Weight	Description	Due Date	Considerations and Late Penalties
Midterm exam	20%	Formal, closed book, written exam that focuses on theory and concepts discussed in the lectures with some calculations of statistical parameters. The midterm covers all topics up to multivariate statistics.	90-minute exam during the lab time of week 6	Missed midterms for valid reasons (see student handbook) can be retaken in week 8.
Final exam	40%	Formal, closed book, written exam that focuses on theory and concepts discussed in the lectures with some calculations of statistical parameters. The final exam covers all topics with emphasis on multivariate statistics.	To be scheduled during the final exam period.	Missed final exams are handled by Service Point.
Data project report	40%	Detailed study of a suite of metamorphic rocks with a pelitic protolith in thin section, hand specimen and through (thermodynamic) calculations.	**Before** Wed, May 1 at 9:00	No late reports will be accepted, because I will be teaching fieldschool from this date

The data project is conducted in groups of 2 or 3 with one report handed in per group. This report is to present the results of a thorough analysis of the dataset using the full diversity of statistical methods discussed in the course. The datasets contain a wealth of statistically interesting features and it is impossible to discover all. That is not the point of the project and I will not grade your report based on whether or not you found everything and tried every technique. The purpose is to dissect and understand the dataset so that you are able to interpret the data in a geological and geochemical context, and your reports will be graded on the level of insight into these data. There are many ways to dissect a dataset and there are generally a variety of statistical techniques that will lead you to the same conclusion. So feel free to attack this dataset in whatever way you like, but the following statistical tools should at least be included;

- data descriptors (*e.g.,* mean, IQR, median, mode etc)
- scatter diagrams, box-and-whiskers plots, histograms
- tests of distribution, cumulative frequency diagrams
- correlation tests and correlation matrices
- t-tests, F-tests or their rank-equivalents

- analysis of variance
- cluster and/or discriminant function analysis
- principle component and/or factor analysis
- (multiple) regression analysis
- spatial analysis of the data, bubble plot maps, semivariograms

The report should be approximately 10 pages in length excluding tables and figures (these are to be put into appendices). Further details on the report will be shared at the start of the project.

Assessment rubrics: This course uses the Faculty of Science assessment rubric, copied below.

Score	**Identification of relevant concepts** (Choice of correct model, theory, equation, etc.);	**Correct application of concepts** (Correct combination and application of models, theories, equations, etc.)	**Efficiency of approach** (No extra steps or extraneous information given)	**Quality of presentation** (Clarity, language, nomenclature, citing specific sources or examples where appropriate)
0	Concepts identified are completely irrelevant, or no concepts identified at all.	Concepts are applied completely incorrectly, or no attempt has been made to apply concepts.	The entirety of the work presented is unnecessary or irrelevant, or no approach has been taken at all.	Work is unclear and fails to use appropriate nomenclature. Citations (where required) are absent.
1	Some identified concepts are at least partly correct. Important concepts are missing and/or incorrect concepts are identified	Some concepts have been combined or applied in a partially appropriate manner. Important steps or syntheses are missing and/or incorrect steps are taken.	Much of the work presented is unnecessary or irrelevant.	Work is largely unclear and only occasionally uses appropriate nomenclature. Citations (where required) are substandard.
2	The identified concepts are largely correct and partly complete.	The application of concepts is somewhat appropriate with multiple minor, or a few major errors.	Some unnecessary steps are taken and/or unnecessary information is given.	Work is partly clear and uses some appropriate nomenclature. Citations (where required) are substandard.
3	The identified concepts are largely correct and mostly complete.	The application of concepts is largely appropriate with no major errors and few minor ones.	Few unnecessary steps are taken and/or unnecessary information is given.	Work is generally clear and uses appropriate nomenclature. Citations (where required) are appropriate.
4	All of the relevant concepts are identified, and no incorrect concepts are chosen.	The application of concepts is entirely correct and error-free.	No unnecessary steps are taken and no unnecessary information is given.	Work is at the highest level of clarity, using entirely appropriate nomenclature. Citations (where required) are comprehensive.

Use of AI tools;

AI and machine learning tools have the potential to transform (statistical) data analysis, making it faster and more effective, and allowing application of more diverse and advanced methods, thereby gaining a deeper understanding of your data. However, AI is susceptible to hallucination and it can be hard to understand how it arrived at its answer. Moreover, results may not be robust or meaningful. AI and machine learning are therefore a valid (and potentially powerful) aid in your data analysis, but you remain responsible for critically evaluating its output. Questions you should ask yourself include: Is the output correct; Is the output meaningful; Are fits and models robust; Do I understand how it arrived at the output/conclusions; Does the analysis meet data and method requirements; Is it the most appropriate method; *etc*. **Generative AI tools are therefore allowed in this course**, including for the assessed Data analysis project, but any results obtained by using such tools **must be attributed** (*e.g.* type of tool and prompts used) and **output has to be critically assessed** (*e.g.* through method evaluation or test-set verification).

General McGill policy statements;

Instructor generated course materials (e.g., handouts, notes, summaries, exam questions, etc.) are protected by law and may not be copied or distributed in any form or in any medium without explicit permission of the instructor. Note that infringements of copyright can be subject to follow up by the University under the Code of Student Conduct and Disciplinary Procedures

Assessments in this course are governed by the Policy on Assessment of Student Learning (PASL), which provides a set of common principles to guide the assessment of students' learning. Also see Faculty of Science-specific rules on the implementation of PASL.

Legally mandated academic accommodations are handled by Student Accessibility and Achievement. For more information see https://www.mcgill.ca/access-achieve/

In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French written work that is to be graded. This does not apply to courses in which acquiring proficiency in a language is one of the objectives." (Approved by Senate on 21 January 2009)

Conformément à la Charte des droits de l'étudiant de l'Université McGill, chaque étudiant a le droit de soumettre en français ou en anglais tout travail écrit devant être noté, sauf dans le cas des cours dont l'un des objets est la maîtrise d'une langue. (Énoncé approuvé par le Sénat le 21 janvier 2009)

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures" (Approved by Senate on 29 January 2003) (See McGill's guide to academic honesty for more information).

In the event of extraordinary circumstances beyond the University's control, the content and/or assessment tasks in this course are subject to change and students will be advised of the change.