IELE756 Preparación y Análisis de Datos

Spring 2025

Leo Ferres, PhD

September 5, 2025

Contents

1 Overview

Curious how environmental exposures (air quality, climate, greenness) and agricultural production patterns shape the regional burden of noncommunicable diseases in Chile? Want to turn official statistics and satellite records into defensible models that illuminate territorial inequalities? Prefer working in teams on a fast, policy-relevant problem rather than passive lectures?

In this project-driven course, you will assemble an annual, region-level dataset for 2010–2022 that links MINSAL hospital discharges, INE crop production, CASEN socioeconomic indicators, ERA5 climate, Sentinel-5P atmospheric composition, VIIRS nightlights (population proxy), Earth Engine greenness indices (NDVI, SAVI, NDWI), Köppen–Geiger climate classes, and Chilean administrative boundaries. You will pose and test hypotheses

about how exposure, agriculture, climate zones, urbanization, and socioeconomic context relate to NCD burden, culminating in hierarchical Bayesian models and uncertainty-aware regional maps designed to inform policy.

2 Specifics

- **Class Time**: Thursdays (11:10am-12:20pm), Fridays (11:10am-1:40pm)
- **Class Room**: Online (but first class is in 116 Q Building, 1st floor)
- **First Class**: Friday, September 6, 2025
- **Office Hours**: by appointment (email staff)
- **Course Website**: https://leoferres.github.io/iele25-2.html

---

- **Professor**: Leo Ferres, 227 S Building, lferres@udd.cl
- **TAs**:
  - Antuan Vayisqui, avayisquia@udd.cl
  - Alex Palatnic, apalatnicf@udd.cl
- **Staff Email**: (send email through Canvas to all staff)

- **Prerequisites**:
  - (Taller de programación en Python)
  - IIP225A (Probability and Statistics)
  - or permission from the instructor

---

3 Inverted Lectures

This is not a typical course and that's by design. Rather than traditional lectures, this is a project-driven, skill-based course designed to immerse you in the actual practice of scientific modeling and data analysis. You won't be sitting through lectures while I walk you through material. Instead, you'll learn by doing, tackling real-world problems using tools and methods drawn from contemporary data science and statistical modeling. Every week (or sometimes every two weeks, depending on the complexity), you'll receive an assignment that introduces a specific type of skill or analytical technique. These are not isolated exercises, they build toward a larger goal: constructing a functioning predictive model of dengue spread in Chile. Think of each assignment as a piece of that puzzle.

The structure is intentionally **inverted**: I won't be lecturing on each topic in advance. Instead, I'll guide you to high-quality external tutorials, documentation, and video content (e.g., from YouTube or official project pages). You'll study the material independently before class. Then, **during class, we'll use our time together to dive deeper, discuss the content, troubleshoot your work, and collaboratively address any roadblocks**. To make this work, your engagement is essential. You're expected to come to class having reviewed the assigned materials, having tried the assignment, and critically having real questions. The learning will happen through interaction, collaboration, and problem-solving. My role is not to lecture, but to help you think, guide you through complexity, and learn alongside you as a team.

This course is designed to feel closer to a research lab or collaborative project environment than a lecture hall. If you're curious, self-directed, and ready to take on challenging, meaningful problems, you'll thrive here.

4 Topics

This is a data-driven course on modeling spatial, environmental, and epidemiological phenomena using Bayesian methods. Emphasis is on integrating diverse datasets and implementing models in Bambi and Python. Topics include:

- Tools for data science workflows
  - Google colab
  - Python: pandas, geopandas, xarray, rioxarray, rasterio, dask

  - Obsidian, GitHub, markdown
- Spatial data
  - Administrative boundaries vs grid systems
  - Projections and zonal statistics
- Population estimation
  - Census vs VIIRS nighttime lights
  - Intercensal interpolation
- Environmental monitoring
  - ERA5 climate variables
  - Sentinel-5P gases (NO2, O3, SO2)
  - Greenness indices: NDVI, SAVI, NDWI
  - Köppen-Geiger climate classes
- Agriculture and land surface
  - INE crop harvests
  - Disaggregating "other" and proportional allocation
- Epidemiological and socio-economic datasets
  - MINSAL NCD outcomes
  - CASEN/Census indicators
  - Administrative geodata
- Bayesian data analysis
  - Introduction to Bayesian inference
  - Prior specification
  - Prior and posterior predictive checks
  - Bayesian workflow
- Regression modeling
  - Counts with offsets: Poisson, negative binomial, zero-inflated
  - Nonlinearities and interactions: splines, greenness $\times$ pollution

- Spatiotemporal modeling
  - Hierarchical models: CAR, BYM2, random walks
  - Distributed lag non-linear models for climate and pollutants

5 Readings and Resources

There is no textbook for this class, but there are several good resources. I will make them available asap.

6 Requirements & Grading

There are six components to the class, some of which are required:

- [Required] Engage with course content: Watch and read all assigned video lectures and materials. Mark each module as completed in Canvas, and submit either a short reflection or a question about the content. You earn the full 10 points by completing at least 90% of assigned modules.
- Class participation: Contribute to the learning community by coauthoring at least one nontrivial Canvas post each week about the topics we're covering. Posts can present a solution you're developing, or a thoughtful question about the new material. A trivial post might say "Use `pandas` to sort." A nontrivial post might compare different approaches, explain a failed attempt, or raise a deeper conceptual issue. This is graded as an individual task.
- Assignments. Assignments will be posted weekly (in Spanish).

The grading scheme (adding up to 70) this semester is:

1. 10 points for watching lectures
2. 10 points for in-class participation
3. 50 points for problem sets (5 problem sets plus Assignment 0)
   - Assignment 0 (installation/setup): 5 points
   - Problem Sets 1-5: 9 points each ($5 \times 9 = 45$ points)
   - Total: 50 points

One additional rule is that you cannot pass the class without watching most of the lectures (according to the fairly minimal requirements above), which is the meaning of "required".

7 Past and Future

A similar class is offered often. It was given in the Fall as IELE756 2025-1. You can also go to my homepage.