IELE756 Preparación y Análisis de Datos

Fall 2026 - Three-Dataset Edition

Leo Ferres, PhD

May 9, 2026

Overview

How did Chile go from roughly 2% foreign-born in 2002 to over 8% in 2024? Where do immigrants settle, what work do they find - and do their health outcomes differ from those of Chilean-born residents?

The **Census 2024 microdatos** give us the most detailed snapshot ever of Chile's population: over 19 million individual records linked to their households and dwellings. Two complementary health datasets enrich the picture:

- **ENO** (Enfermedades de Notificación Obligatoria) - the national mandatory notifiable-disease surveillance registry, with ~333,000 records from 2007–2024.
- **GRD** (Grupos Relacionados de Diagnóstico) - hospital discharge records covering ~5 million+ episodes from 2019–2024, coded with ICD-10 diagnoses.

In this project-driven course you will weave all three datasets into a coherent narrative: **Migration, Health, and Socioeconomic Integration in Chile**. Working in pairs, each team will be assigned a set of **1 to 3 comunas** (mostly from the Región Metropolitana), chosen so that each team covers a roughly similar total population. Each team will build a complete analytical pipeline: from raw parquet/CSV/pipe-delimited files through demographic profiling, disease surveillance analysis, hospital discharge analysis, spatial mapping, and finally cross-dataset ecological modeling at the **comuna level**.

Because the three datasets contain *different people* (a census respondent is not the same individual as an ENO notification or a hospital discharge), cross-dataset analysis is necessarily **ecological**: you will aggregate each dataset to the comuna level and link them by `codigo_comuna`. Understanding when ecological inference is valid - and when it is not - is an explicit learning objective.

The final deliverable is a **video (8–12 minutes)** presenting your findings as an integrated research story, backed by a **GitHub repository** with all code, data

pipeline, and reproduction instructions.

Specifics

- **Class Time**: Thursdays 12:30–13:40, Fridays 11:10–13:40
- **Modality**: Online / Streaming
- **First Class**: Thursday, March 5, 2026
- **Trimester**: March 2 – April 30, 2026
- **Certamen week (no class)**: April 13–18
- **Office Hours**: by appointment (email staff)
- **Course Website**: https://leoferres.github.io/iele26-1.html

---

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

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

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 data analysis and statistical modeling. You won't be sitting through lectures while I walk you through material. Instead, you'll learn by doing, tackling real census microdata using tools and methods drawn from contemporary data science.

Every week you'll receive an assignment that introduces a specific skill or analytical technique. These are not isolated exercises: they build toward a larger goal: constructing a complete analytical portrait of migration and integration in your assigned comunas. Think of each assignment as a piece of that puzzle.

The structure is intentionally *inverted*: I won't lecture on each topic in advance. Instead, I'll point you to high-quality tutorials, documentation, and video content. 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 roadblocks. Your engagement is essential: come to class having reviewed the assigned materials, having tried the assignment, and critically having real questions.

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.

Topics

This is a data-driven course on analyzing census and health microdata with a focus on migration, health outcomes, and socioeconomic integration. Emphasis is on building reproducible analytical pipelines in Python. Topics include:

Tools for data science workflows

- Jupyter notebooks (Google Colab, VS Code, JupyterLab, or any IDE), GitHub, Markdown
- Python: pandas, pyarrow, geopandas, matplotlib, seaborn, plotly
- statsmodels for regression

Census 2024 microdata

- Three tables: vivienda, hogar, persona
- Linking keys: `id_vivienda`, `id_hogar`, `id_persona`
- Parquet files, filtering, joining hierarchical tables
- Handling missing values (`-99`, `NA`), calculated variables

ENO - Notifiable disease surveillance

- Semicolon-delimited CSV, 2007–2024
- Key fields: disease code, notification date, comuna of residence, nationality, education
- Cleaning nationality coding ("Desconocido" category)
- Notification rates over time, disease profiles by nationality

GRD - Hospital discharge records

- Pipe-delimited files (one per year, zipped), 2019–2024
- Key fields: `DIAGNOSTICO1`, `COMUNA`, nationality, length of stay, severity
- ICD-10 lookup via `CIE-10.xlsx`; diagnostic chapter grouping
- Loading only needed columns, filtering to your comunas immediately (large files)

Demographic analysis

- Age pyramids by sex, dependency ratios
- Household typology, household size distributions

Migration variables

- Place of residence 5 years ago (`p24_lug_resid5`)
- Place of birth (`p25_lug_nacimiento`)
- Arrival period (`p26_llegada_periodo`)
- Nationality (`p27_nacionalidad`)

Labor force participation

- Labor force status (`sit_fuerza_trabajo`)
- Occupation (`cod_ciuo`), economic activity (`cod_caenes`)
- Commute mode (`p45_medio_transporte`)

Education

- Years of schooling (`escolaridad`), education level (`cine11`)
- School attendance variables (`asistencia_*`)

Housing and services

- Housing quality (`materialidad`), overcrowding (`hacinamiento`)
- Internet access, tenure (`tenencia`)

Spatial analysis

- Mapping at comuna level with geopandas + shapefiles
- Choropleth maps, spatial distributions

Exploratory data analysis

- Distributions, cross-tabulations, correlations
- Grouped bar charts, heatmaps, scatter plots

Statistical modeling

- Logistic regression for binary census outcomes
- Count models: Poisson and Negative Binomial regression (for notification/discharge counts)
- Ecological regression: linking aggregate census predictors to health outcomes at the comuna level
- Model diagnostics, coefficient interpretation (odds ratios, incidence rate ratios)
- Predicted probability/rate maps

Weekly Schedule

Week	Dates	Topic	Assignment
0*	Mar 5–6	Intro + tools; overview of all three datasets	Tarea 0 released
1	Mar 12–13	Census data structure, parquet, pandas + pyarrow	Tarea 0 due; Tarea 1 released

Week	Dates	Topic	Assignment
2	Mar 19–20	Census: joins, filtering, demographic + migration indicators	
3	Mar 26–27	ENO + GRD: loading, cleaning, ICD-10 lookups, disease profiles	Tarea 1 due; Tarea 2 released
4	Apr 2–3	Health data deep dive: rates, age groups, spatial mapping	
5	Apr 9–10	*Certamen week: no class*	
6	Apr 16–17	Cross-dataset ecological analysis at comuna level	Tarea 2 due; Tarea 3 released
7	Apr 23–24	Modeling + integration: ecological regression, synthesis	
8	Apr 30	Tarea 3 wrap-up; Final project released	Tarea 3 due; Final project released
9	May 15	(no class) Anomaly-proposal gate	Anomaly proposal due (ungraded)
10	May 29	(no class) Repo + video deadline	Final repo + video due
11*	Jun 5	**In-person**: handwritten memo (individual)	Final memo

- The star (*) means that particular class is in person. # Readings and Resources

There is no textbook for this class, but the following resources are essential:

Census 2024 documentation

- Census manual: `materials/census2024/manual_uso_microdatos_censo2024.pdf`
- Variable dictionary: `materials/census2024/diccionario_variables_censo2024.xlsx`
- Variable glosses: `materials/census2024/diccionario_variables_glosas_censo2024.xlsx`
- Data download: https://censo2024.ine.gob.cl/resultados/

ENO documentation

- Variable dictionary: `materials/eno/diccionario_de_variables_eno_2007_2022fin.xlsx`
- Methodology: `materials/eno/metodologia_validacion_anonimizacion_y_publicacion_bases_de_d`
- Data: `materials/eno/20241218_base_eno_final.csv`

GRD documentation

- ICD-10 lookup table: `materials/grd/CIE-10.xlsx`
- Master tables (diagnoses, procedures, etc.): `materials/grd/TablasMaestrasBasesGRD.xlsx`
- Data (zipped, one file per year): `materials/grd/GRD_PUBLICO_20{19..24}.zip`

ICD-10 reference

- WHO ICD-10 online browser: https://icd.who.int/browse10/2019/en

Python tutorials

- pandas
- geopandas
- matplotlib
- Google Colab (in Spanish; also works with JupyterLab, VS Code, or any Jupyter-compatible IDE)

Reference

- R for Data Science (Spanish): https://es.r4ds.hadley.nz/ (es muy bueno, pero en R)
- Working with folders and files (Windows); from the console

Assignments

All work is done in **pairs** (parejas), each assigned a set of 1 to 3 comunas (mostly from RM) with roughly equal total population. Detailed problem set descriptions will be posted as separate files; what follows is a summary.

The three datasets are introduced **progressively**: Tarea 0 gives a shallow first contact with all three; Tarea 1 goes deep on Census; Tarea 2 goes deep on ENO + GRD; and Tarea 3 merges everything at the comuna level.

Tarea 0: Setup & First Contact with All Three Datasets (5 pts)

Install tools (Jupyter-compatible IDE, GitHub account, Python environment). Create a GitHub repo for your team. Then load each dataset for your assigned comunas and report basic shape/info:

- **Census**: load the parquet files, filter to your assigned comunas using pandas via pyarrow, report shape/info, show first rows.
- **ENO**: load the CSV (semicolon-delimited), filter to your comunas, count notifications by year, show the top 5 diseases.
- **GRD**: unzip one year, load the pipe-delimited file, filter to your comunas, join `DIAGNOSTICO1` to the CIE-10 lookup table, show top 5 diagnoses.

Due: Week 1.

Tarea 1: Demographic Profile & Migration Landscape (10 pts)

Build a demographic and migration portrait of your assigned comunas using the Census persona table.

- Load and join vivienda + hogar + persona tables
- Age pyramids by sex, overlaid Chilean-born vs. foreign-born

- % foreign-born by comuna, top nationalities, migration status (`p24_lug_resid5`)
- Choropleth maps at comuna level (population, % immigrants), bar charts of nationality distributions
- Compare your comunas to national/regional averages
- **Key output**: a **comuna-level summary table** (`codigo_comuna`, population, % foreign-born, median age, mean years of education, employment rate - by nationality group: Chilean, Foreign). This table will be reused in Tarea 3.
- **Libraries**: pandas, pyarrow, geopandas, matplotlib
- **Due: end of Week 3**

Tarea 2: Health Landscape - ENO + GRD (10 pts)

Part A - ENO (Notifiable Diseases)

- Filter ENO to your comunas, clean nationality/education coding (report and exclude "Desconocido" from nationality-specific rates)
- Notification rates over time for top diseases
- Disease profiles by nationality (Chilean vs. Foreign): over/under-representation
- Choropleth of notification rates by comuna
- **Output**: a **comuna-level ENO summary table** (notification counts and rates by disease group and nationality)

Part B - GRD (Hospital Discharges)

- Load 2022–2024 GRD files, filter to your comunas
- Join diagnoses to CIE-10 names
- Average length of stay, top diagnostic chapters, severity distribution - all by nationality
- Hospitalization rate map by comuna
- **Output**: a **comuna-level GRD summary table** (discharges, average length of stay, severity - by nationality)

Libraries: pandas, geopandas, matplotlib, seaborn or plotly

Due: end of Week 6

Tarea 3: Cross-Dataset Ecological Modeling (10 pts)

- **Merge** the three comuna-level summary tables from Tarea 1 and Tarea 2 on `codigo_comuna`
- Scatter plots, correlation matrices: which census variables predict which health outcomes?
- **Model**: Poisson or Negative Binomial regression for notification counts (with population offset), or linear/logistic for severity/length of stay

- Coefficient interpretation (incidence rate ratios, odds ratios), explicit discussion of the **ecological fallacy**
- Visualization: coefficient plots, predicted rate maps, residual maps
- **Libraries**: statsmodels, geopandas, matplotlib
- **Due: end of Week 8**

Final Project: One Anomaly, Defended (25 pts)

By the end of Tarea 3 your pipeline has produced thousands of numbers. The final project asks you to pick **one** finding that genuinely surprised you and defend it well enough that an outsider believes you understand it. The full specification is in `assignments/FinalProject.md`. Three components:

1. **GitHub repository** (10 pts, async): the full pipeline plus a `final_anomaly.ipynb` that isolates the chosen anomaly, with a README that includes an AI-use disclosure.
2. **Video** (7 pts, async, 8 to 10 minutes): an anomaly-centric defense, not a survey of the three datasets. Both partners speak.
3. **Handwritten in-person memo** (8 pts, individual, 45 min): no laptops, no AI, no notes. Each partner writes their own.

A non-graded **anomaly proposal** is required as a mid-process gate (due about a week after Tarea 3). No proposal, no project.

AI policy is split: async components allow AI with disclosure; the in-person memo forbids it. The memo is 8 of 25 points and is designed to verify that the async work represents real understanding.

Requirements & Grading

The grading scheme adds up to **70 points**:

Component	**Points**
Participation (engagement, questions, collaboration)	10
Tarea 0 (setup + first contact)	5
Tarea 1 (demographics + migration)	10
Tarea 2 (health landscape: ENO + GRD)	10
Tarea 3 (ecological modeling)	10
Final project: GitHub repo	10
Final project: Video	7
Final project: Handwritten memo (individual)	8
Total	**70**

One additional rule: you cannot pass the class without completing most of the core assignments and participating actively, as defined above.

Final Project Rubric

Detailed rubrics for each component live in `assignments/FinalProject.md`. Summary:

GitHub Repo (10 pts)

Criterion	Pts	What to check
Reproducibility (clone + run reproduces the headline figure)	3	`requirements.txt` works, README is followable in 2 min
Pipeline correctness (joins, offsets, units verified by spot-check)	3	Computed values match documentation; no silent imputation
`final_anomaly.ipynb` quality (anomaly isolated, alternatives explored in code)	2	Two or more alternative-explanation checks present
README + organization + AI-use disclosure	2	Clear, honest, complete

Video (7 pts)

Criterion	Pts	What to check
Anomaly framing (specific, falsifiable, grounded in numbers)	2	Not vibes
Depth of explanation (alternatives ruled out with evidence)	3	Evidence, not assertion
Visual + delivery quality (legible figures, both partners speak)	2	Watchable, paced for an outside viewer

Handwritten in-person memo (8 pts, individual)

In person, on paper, no AI / laptops / notes. 45 minutes, three prompts drawn on the day from the bank in `FinalProject.md`. Graded per student, not per pair.

Criterion	Pts	What to check
Specificity (real numbers, comunas, code paths from memory)	3	Cited, not hand-waved
Critical reasoning (alternatives, falsification, limits)	3	Reasoning, not assertion
Coherence with team artefacts	2	What you wrote matches the repo and video

Policies

- **Late policy**: 2-day extension available by email to staff before the deadline.
- **Collaboration**: Work is done in pairs. Both members must understand all code submitted.
- **AI policy**:
  - *Tareas 0 to 3 and the final project's async components (repo + video)*: AI tools are allowed. Disclose them in the README of your final repo.
  - *Final project's in-person memo*: no AI, no laptops, no notes. This is the contract that gives the async work its meaning.
- **Submission**: Canvas (PDF of notebook + repo link for the final project; video link; the memo happens live).

Past and Future

A similar class is offered often. It was given in the Fall of 2025 and the Spring 2025. You can also visit my homepage and check out my blog at https://leoferres.blog, which is updated more frequently.

Typeset with: `pandoc iele26-1-3datasets.md -o iele26-1-3datasets.pdf --pdf-engine=pdflatex && evince iele26-1-3datasets.pdf &`