Tarea 3: Cross-Dataset Ecological Modeling

IELE756 – Preparación y Análisis de Datos

Leo Ferres, PhD

April 16, 2026

Tarea 3: Cross-Dataset Ecological Modeling

Points: 10
Released: Thursday, April 16, 2026
Due: Thursday, April 30, 2026 (before class)
Submission: Canvas – PDF export of your notebook + link to your GitHub repo

---

Goal

This is the final and most demanding analytical assignment of the course. You now have three comuna-level summary tables: one from the Census (Tarea 1), one from ENO, and one from GRD (both Tarea 2). Your job is to **link them into a single analytical dataset** and ask whether demographic composition predicts health outcomes at the comuna level. You will fit count-regression and continuous-outcome models, interpret their coefficients, produce predicted-rate maps, and write a careful discussion of the **ecological fallacy**.

Unlike Tareas 1 and 2, this assignment gives you far less scaffolding. You have been working with these data for six weeks. You know how to load a CSV, how to merge on a key, and how to make a bar chart. The instructions below tell you **what** to produce and **which tools** to reach for; the **how** is on you.

By the end of this assignment you should be comfortable with:

- Assembling a multi-source analytical table from heterogeneous pipelines
- Fitting Poisson and Negative Binomial regression with offsets
- Fitting linear (or logistic) regression for continuous or binary comuna-level outcomes
- Reading regression output: coefficients, standard errors, p-values, incidence-rate ratios, model fit diagnostics

- Articulating the difference between an individual-level causal claim and an ecological association
- Turning a statistical result into a map and a paragraph

Part 0: Building the Analytical Table (1 pt)

0.1 Class-pooled master table

Your own team has at most three comunas. Three rows is not a dataset. For Tarea 3 we **pool across the class**: every team published their Tarea 1 and Tarea 2 summary tables as **Quiz 1**, using a fixed schema and a fixed file-naming convention. Open the Canvas group **"IELE756 Tarea 3 data pool"** (via **People** -> **View User Groups**), go to its **Files** tab, select all entries, and download them as a single ZIP. Extract into one folder (for example, `shared/`). You should end up with 63 CSVs: 21 census, 21 ENO, 21 GRD.

Concatenate each family into a single master DataFrame (pandas `concat` over the result of `glob.glob("shared/census_team*.csv")` is enough). You should end up with three master tables, one row per RM comuna.

Run the following sanity checks and report the output in your notebook:

- Row count per master table, and number of unique `codigo_comuna` values. These should agree; if they do not, you have duplicates from overlapping team assignments (investigate).
- Number of comunas that appear in the census master but not in ENO or GRD (and vice versa). Document any asymmetries.
- Locate your own assigned comunas in each master and verify that the numbers match the summary tables you produced in Tareas 1 and 2. If they do not match, your own Quiz 1 upload was likely wrong; state in a Markdown cell what you found and which version you are using going forward.

If a classmate's CSV is missing or malformed, ping them on the course channel and move on with a reduced N; do not silently impute missing rows. State the final N you use for all downstream models.

0.2 Merge on `codigo_comuna`

Merge the three master tables into a single wide DataFrame keyed by `codigo_comuna`. Think carefully about merge type: you want one row per comuna, and you need to handle comunas that appear in one dataset but not another (for example, comunas where every ENO row was anonymized).

Report the final shape and list any comunas that were lost at each merge step. Do not silently drop rows.

0.3 Derived variables

Create, at minimum, the following comuna-level covariates from the Census columns:

- `pct_foreign` (already in the Census summary)
- `log_pop_total` (natural log of total population)
- `pct_unemployed` (1 minus your combined employment rate, or derive it directly from `sit_fuerza_trabajo` if you have unemployment counts)
- `schooling_gap` = `mean_schooling_chilean` minus `mean_schooling_foreign`
- Any one additional covariate you justify in a Markdown cell (examples: % aged 65+, % in overcrowded housing, % without internet access)

If a covariate you want is not in the master summary but is in the Census microdata, recompute it at the comuna level and add it as a new column. Document every new variable.

Part 1: Exploratory Cross-Dataset Analysis (1.5 pts)

1.1 Correlation matrix

Compute a Pearson correlation matrix covering your census covariates and your two primary health outcomes: `eno_rate_per_10k` and `grd_rate_per_10k`. Display it as a heatmap.

Include at least six covariates. Order the matrix so that related variables cluster visually. Comment in Markdown on the three strongest correlations and whether their signs match your expectations.

1.2 Bivariate scatter plots

Produce a small multiple of scatter plots: each panel shows one census covariate on the x-axis and one health outcome on the y-axis. Overlay a simple OLS fit line (seaborn's `regplot` or `lmplot` is enough). Label each point with its comuna name for the five comunas with the most extreme residuals.

1.3 Outlier and leverage discussion

Identify at least two comunas that behave as visual outliers on the scatter plots. In one short paragraph each, propose a plausible substantive explanation (a hospital that draws from outside the comuna? an unusually old population? an anonymized ENO burden?). You are not expected to prove these hypotheses, only to articulate them clearly.

Part 2: Count-Data Regression (2.5 pts)

The ENO and GRD outcomes are **counts per comuna per time window**. Rates depend on population. You will model them with a **log-linear count model** and a **population offset**.

2.1 Poisson regression (0.75 pts)

Fit a Poisson regression of `eno_total` (count of ENO notifications per comuna) on your covariates, using `log(pop_total)` as an offset. Use `statsmodels.api.GLM` with `family=sm.families.Poisson()`, or the equivalent formula interface.

Required covariates: at least `pct_foreign`, one education measure, one employment measure, and one housing/service measure.

Report:

- The full coefficient table (estimates, standard errors, z-values, p-values)
- A column of **incidence rate ratios** (IRR = `exp(coef)`) with their 95% confidence intervals
- Model fit: deviance, Pearson chi-square, AIC

2.2 Check for overdispersion (0.5 pts)

Poisson regression assumes variance equals the mean. Compute the dispersion statistic (Pearson chi-square divided by residual degrees of freedom). If it is substantially greater than 1, the data are **overdispersed** and Poisson standard errors are under-stated.

Briefly report your dispersion value and what it implies.

2.3 Negative Binomial regression (0.75 pts)

Refit the same model using Negative Binomial regression (statsmodels `NegativeBinomial` or `GLM` with `NegativeBinomial` family). Show the coefficient table side-by-side with the Poisson one. Point out any coefficient whose p-value changes in a way that would alter your conclusions. Recommend which model you would report as your primary result, and justify the choice.

2.4 Repeat for GRD (0.5 pts)

Fit a second count model using `grd_total` as the outcome, same covariates, same offset. You may use Poisson or Negative Binomial, justified by your dispersion check. Present the IRR table and one or two sentences interpreting the largest effect.

Part 3: Continuous-Outcome Regression (1.5 pts)

3.1 Pick an outcome

Choose **one** continuous comuna-level outcome from the GRD summary, for example:

- `grd_mean_los` (mean length of stay in days)
- `grd_mean_severity`
- `grd_mortality_rate` (treat as a proportion; logit transform optional)

State which you picked and why it is interesting.

3.2 Linear regression

Fit a linear regression (OLS) of your chosen outcome on the same covariates you used in Part 2. Report the coefficient table, R^2, and adjusted R^2.

3.3 Diagnostics

Produce at least two diagnostic plots: residuals versus fitted values, and a QQ plot of residuals. Comment on whether the OLS assumptions look defensible. If they do not, state what you would try next (a transformation? a robust estimator? a different outcome?). You do not need to implement the alternative: a clear diagnosis is enough.

Part 4: Ecological Fallacy (1 pt)

Everything above is **ecological**: the unit of observation is a comuna, not a person. An association between `pct_foreign` and `eno_rate_per_10k` at the comuna level does **not** imply that a foreign-born individual has a higher probability of being notified. The fallacy of assuming the individual-level claim from the ecological one has a name (the **ecological fallacy**, Robinson 1950) and a long history of misuse.

In a single dedicated Markdown section (roughly 400 to 700 words), address the following:

1. State, in your own words, the difference between an individual-level and an ecological association.
2. Give **one concrete example** from your results where an ecological coefficient could be misread as an individual-level causal claim, and explain exactly what the misreading would be.
3. Give **one example** from your results where the ecological association is still useful for public-health planning, even though it cannot support individual-level claims.

4. Briefly discuss at least one additional threat to inference that you have **not** addressed in your models (spatial autocorrelation, omitted confounding, measurement error in the anonymized ENO, small-N sparsity, etc.) and how you might address it in future work.

No new code is required in this part. Prose quality counts.

Part 5: Spatial Visualization of Model Output (1 pt)

5.1 Predicted-rate map

Using your preferred count model from Part 2 (Poisson or Negative Binomial), compute the predicted notification rate per 10,000 for each comuna and produce a choropleth. Use a diverging or sequential palette as appropriate, and include a clear legend and title.

5.2 Residual map

Compute standardized residuals (Pearson residuals) from the same model and plot them on the same comuna geometry. Use a diverging palette centred at zero. Which comunas have the largest positive residuals? Which have the largest negative? One or two sentences of interpretation.

5.3 Coefficient plot

Produce a **coefficient plot** (also called a forest plot) for your primary count model: one row per covariate, with the point estimate and a horizontal 95% confidence interval, log-scale on the x-axis if you use IRR. Coefficient plots should be labelled legibly and should mark the null (IRR = 1 or coefficient = 0).

Part 6: Integrated Synthesis (1.5 pts)

This is the part that ties Tareas 1, 2, and 3 together and that will feed directly into your final video and README.

In no more than **800 words**, write an **integrated findings section** that answers the following three questions:

1. What does the demographic portrait of the Región Metropolitana (Tarea 1) look like at the comuna level? What is the dominant axis of variation across comunas?

2. How does the health landscape (Tarea 2) map onto that demographic axis? Where are the matches, and where are the mismatches?
3. What do the cross-dataset models (Tarea 3) add beyond what is visible in a correlation matrix? Are there findings that would not have been obvious from any single dataset?

Keep it tight. Reference specific numbers from your tables and specific figures from your notebook. This section will score higher if it is a coherent narrative than if it is a list of bullet points.

Deliverables

Submit on **Canvas** before class on Thursday, April 30:

1. **PDF export** of your notebook.
2. **Link to your GitHub repository**. The repository must contain, at minimum:
   - The notebook for Tarea 3
   - The three master CSVs used as input
   - The merged analytical table, saved to `output/tarea3_analytical_table.csv`
   - A `requirements.txt` or `environment.yml` listing `pandas`, `numpy`, `statsmodels`, `geopandas`, `matplotlib`, `seaborn`

Your notebook must include:

- Markdown explanations for every numbered part
- Executed code cells with visible output
- At minimum the following figures:
  - Correlation heatmap (Part 1.1)
  - Scatter small-multiple with labels (Part 1.2)
  - Coefficient plot (Part 5.3)
  - Predicted-rate choropleth (Part 5.1)
  - Residual choropleth (Part 5.2)
- The full coefficient tables from Parts 2 and 3
- The written ecological-fallacy section (Part 4)
- The integrated synthesis (Part 6)

What We Are Looking For

This assignment is graded less on "did you run the code" and more on "did you think about it". A clean Poisson fit with a muddled discussion will score lower than a Poisson fit with overdispersion correctly diagnosed, reported, and fixed.

Specifically, the grading will weigh:

- **Correctness of the pipeline**: does your merged table agree with the inputs? Are offsets and units right?
- **Appropriateness of model choice**: did you justify Poisson vs. Negative Binomial? Did you check diagnostics before reporting OLS?
- **Quality of interpretation**: do you translate IRRs into sentences a public-health reader could use? Do you distinguish what your data can and cannot say?
- **Figure craft**: legible axes, labelled extremes, appropriate palettes, clear legends.
- **Writing**: tight, specific, referenced. Avoid filler.

Tips and Common Pitfalls

- **Offsets, not predictors**: population enters a count model as an offset (`offset=log(pop_total)`), not as a right-hand-side variable. Using it as a predictor is a common mistake and will distort every other coefficient.
- **Units**: pick one time window per outcome and stick to it. ENO in your summary is 2007–2024 (18 years); GRD is 2022–2024 (3 years). Do not compare the raw counts without acknowledging the denominator mismatch.
- **Collinearity**: `pct_foreign`, `mean_schooling_foreign`, and several other covariates are highly correlated at the comuna level. Check variance inflation factors (`statsmodels` provides `variance_inflation_factor`) and consider dropping the worst offender before reporting.
- **Small-N modesty**: even pooled across the class you have on the order of 50 comunas. Treat p-values as descriptive summaries, not as gates. Confidence intervals are more honest.
- **Do not reinvent**: the Quiz 1 pool is your canonical input. Do not recompute the per-comuna numbers from the original microdata. If a column you need is missing, add it yourself at the **master** level in your notebook, with documentation, rather than editing any team's uploaded CSV.
- **Version your code**: commit early, commit often. The final repository is 15 points of the course; do not leave the committing for April 30.

Grading Breakdown

Part	Points
Part 0: Analytical table assembly	1
Part 1: Exploratory cross-dataset analysis	1.5
Part 2: Count-data regression	2.5

Part	Points
Part 3: Continuous-outcome regression	1.5
Part 4: Ecological fallacy	1
Part 5: Spatial visualization of model output	1
Part 6: Integrated synthesis	1.5
Total	**10**

As in previous tareas, half of each part's score comes from correct code and output; the other half comes from clear Markdown explanations and thoughtful interpretation.

Typeset with: `pandoc assignments/Tarea3.md -o assignments/Tarea3.pdf --pdf-engine=pdflatex && evince assignments/Tarea3.pdf &`