A Reproducible Paper[*]

Using `pixi` and `quarto` and codespaces to handle environments and execution

Elijah Knaap   Another Person

2024-12-06

This project shows how to generate a reproducible environment and execute an entire analysis (including building the paper) via github codespaces.

Introduction

Sometimes people think I know what i'm doing.

Note for this to be visible, you need to install the quarto extension inside the paper directory with

```
quarto install extension sellorm/quarto-social-embeds
```

Literature Review

You should be using open-source. The practices here are a good way to do 'full-stack' open and reproducible science.

Methods

We often say that *spatial is special* because of two phenomena known as spatial dependence and spatial heterogeneity (Anselin 1989, 1988). Two classic spatial econometric models designed to handle these effects include the *Spatial Lag Model*, defined as

[*] This project is supported by the National Science Foundation Grant #2345820: An Open Source Ecosystem for Spatial Data Science.

$$y = \rho W y + \beta X + \epsilon$$

and the *Spatial Error Model* defined as

$$\begin{split} y &= \beta X + u \\ u &= \lambda W u + \epsilon \end{split}$$

The real point being that you can dropdown to real latex any time you need to. If necessary, add the library in the `header-includes` section of the quarto yaml. For example if you needed aligned equations, you can `split` with ampersands where you want to set the alignment

Results

This section uses `quarto`'s conditional formatting to swap out the correct table depending out output. The problem here is `pandas` can write nice latex tables, but those don't convert to html. Instead you can just write both formats out to file and select the correct one on-demand.

Table 1: Blockgroups in San Diego

	n-total-pop	median-household-income
0	1577.000000	150688.000000
1	1673.000000	127292.000000
2	1915.000000	90673.000000
3	1271.000000	65219.000000
4	695.000000	NaN
5	2617.000000	81250.000000
6	500.000000	64631.000000
7	808.000000	64787.000000
8	1682.000000	59010.000000
9	1151.000000	79725.000000

You can also do the same thing with figures, e.g. to swap in an interactive map in the html output and use a static map in the pdf.

Everyone from the `R` world will recognize Figure 1 as coming from `ggplot`. It shows up either way. But the blockgroups in San Diego show up differently depending on the output.

Here is a map showing the spatial graph of blockgroups in San Diego using a Rook contiguity rule. But Figure 2 shows up differently depending on the output, thanks to quarto's conditional include syntax (awesome).

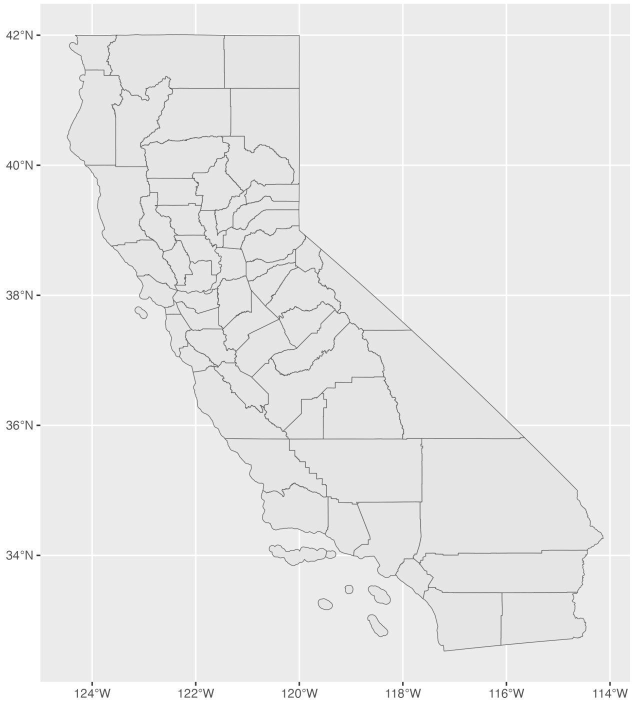

Figure 1: California Counties

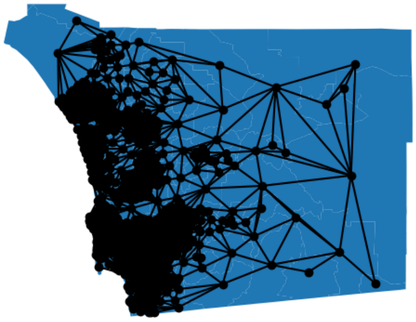

Figure 2: Rook Contiguity Graph for SD Blockgroups

> ℹ Note
>
> This is kinda hacky because it relies on an iFrame that requires the embedded map to be available the relative URL set above (you cant download this html file and expect it to work). The `embed-resources` option in quarto wont work for an iframe either.

In the end, we seamlessly weave together `geosnap` (Knaap and Rey 2024), PySAL (Rey et al. 2021), and `tidycensus` (Walker and Herman 2024). Use the best tool for the job, then let quarto sew things together.

Discussion

Note that since `pixi` also installs binary dependencies, the install is complete with tricky deps like GDAL and the correct GCC compiler (hello windows friends!). That lets you R people use `install.packages` for anything that's not (yet) on conda and your build system is guaranteed to work (though you wont have a lockfile for those packages).

Thus, you get a sandboxed install for this project, with *all deps on all platforms*, and dependencies *right inside the repo*. Just like `npm` (if that's your kinda thing). This is especially nice for making sure your analysis environment isn't broken by the evolving ecosystem while your paper is under review.

Conclusion

This is the way.

References

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. *Operational Regional Science Series*. Vol. 4. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-7799-1.

———. 1989. "What Is Special about Spatial Data?: Alternative Perspectives on Spatial Data Analysis." *Santa Barbara, CA, NCGIA Report*, 84–89. https://escholarship.org/uc/item/3ph5k0d4.

Knaap, Elijah, and Sergio Rey. 2024. "Geosnap: The Geospatial Neighborhood Analysis Package: Open Tools for Urban, Regional, and Neighborhood Science." In. Tacoma, Washington. https://doi.org/10.25080/FVWM4182.

Rey, Sergio J., Luc Anselin, Pedro Amaral, Dani Arribas-Bel, Renan Xavier Cortes, James David Gaboardi, Wei Kang, et al. 2021. "The PySAL Ecosystem: Philosophy and Implementation." *Geographical Analysis*, June, gean.12276. https://doi.org/10.1111/gean.12276.

Walker, Kyle, and Matt Herman. 2024. *Tidycensus: Load US Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames*. https://walker-data.com/tidycensus/.