

API-231 / GIS-PubPol

Meeting 21 (Troubleshooting Session)

Yuri M. Zhukov
Visiting Associate Professor of Public Policy
Harvard Kennedy School

April 16, 2024

Plan for today

0. Plan for next two weeks
1. General data questions
2. General methods questions
3. Project-specific questions
 - a) Please raise your hand, and we'll put you in the queue

General data questions

Migration data

“Where can I find global, annual emigration data by country?”

Migration data (global coverage)

Source/link	Type	Spatial scale	Frequency	Availability
WB Global Bilateral Migration	Migration flows (origin-dest.)	Country	Annual	1960-2000
WB Open Data	Net migration, migrant stock	Country	Annual	1960-2023
IOM Migration Data Portal	Multiple indicators	Country	Annual	1990-2020
Our World in Data	Multiple indicators	Country	Annual	1960-2021
IDMC Data Portal	IDPs from conflict, disasters	Country	Annual	2018-2022

Migration data (sub-national and specialized data)

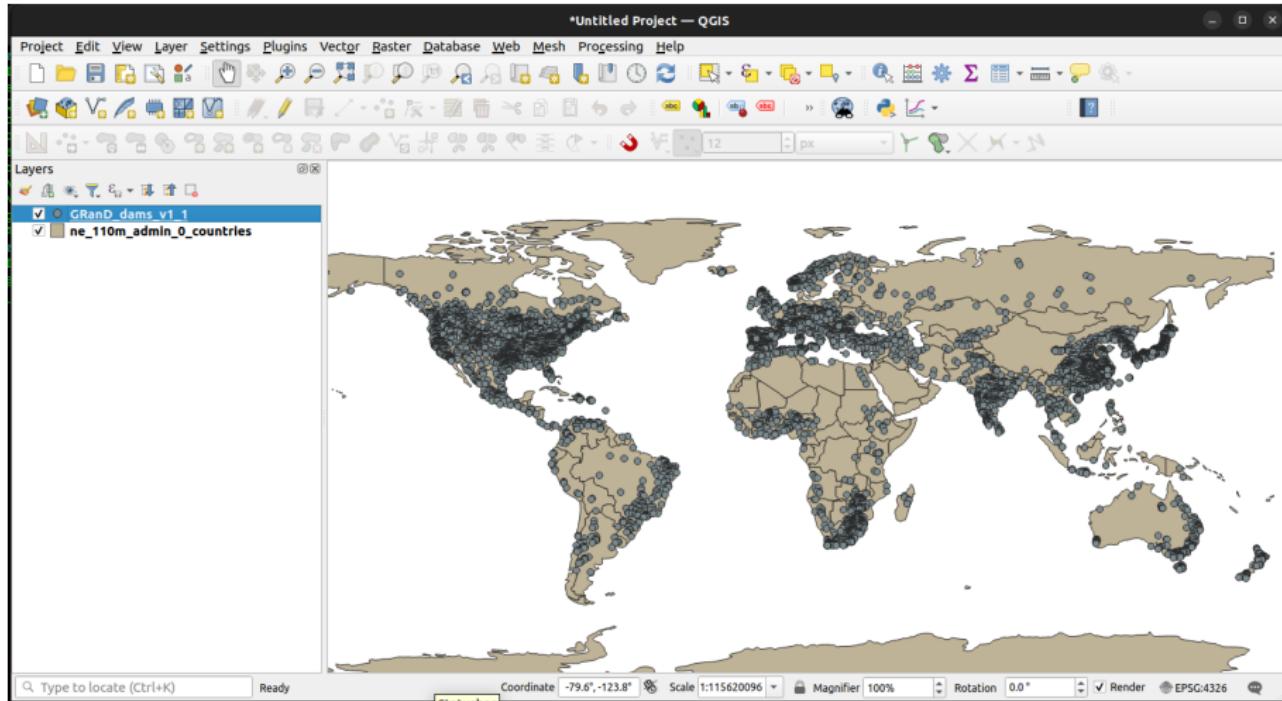
Source/link	Type	Spatial scale	Frequency	Availability
IOM Displacement Tracking Matrix	IDP flows (origin-dest.)	Country, Adm1, Adm2	Variable	2010-2024
UNHCR Data Portal	IDPs, refugees	Country, Adm1	Variable	Variable
CTDC	Human trafficking	Individual	Annual	1960-2023
DHS Immigration Data	Multiple indicators	Points of entry	Monthly	2002-2024

General methods questions

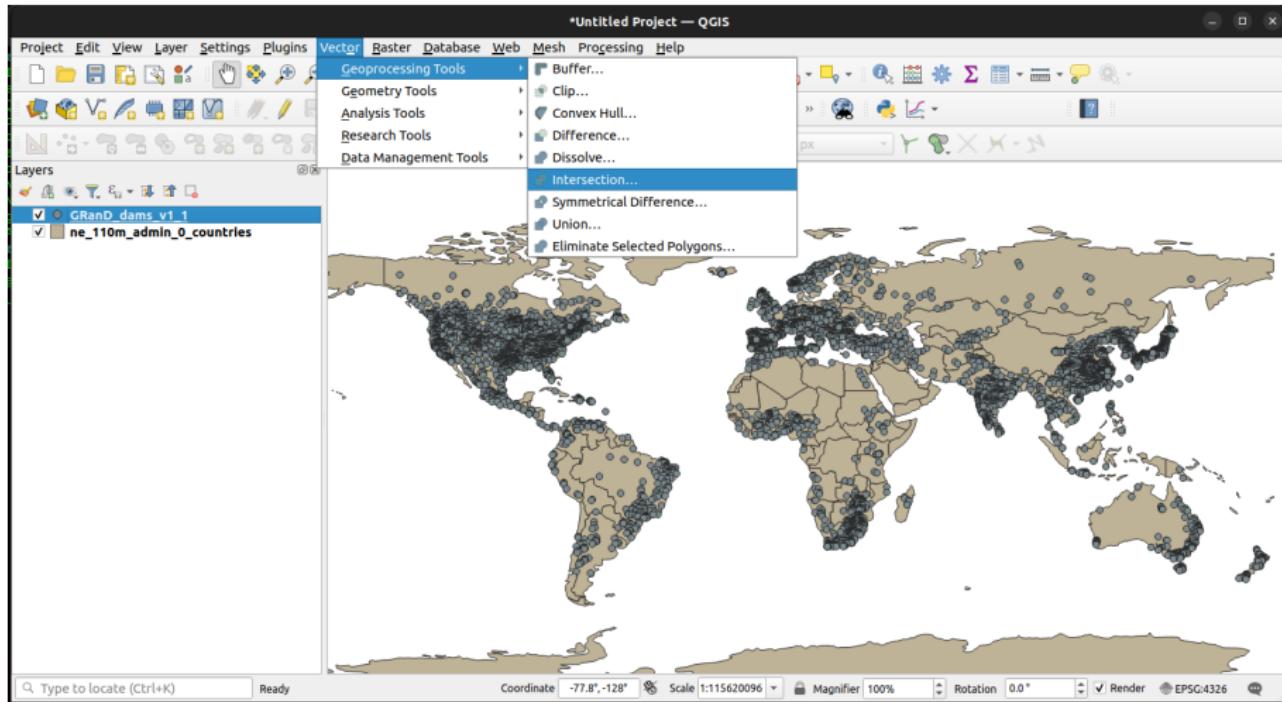
Classifying points by location

"I have a dataset in .CSV format with 700+ rows. I want to classify each data point into one of two categories, based on their geographical location . . . (e.g. classifying whether an oil spill occurred in an offshore or onshore area)."

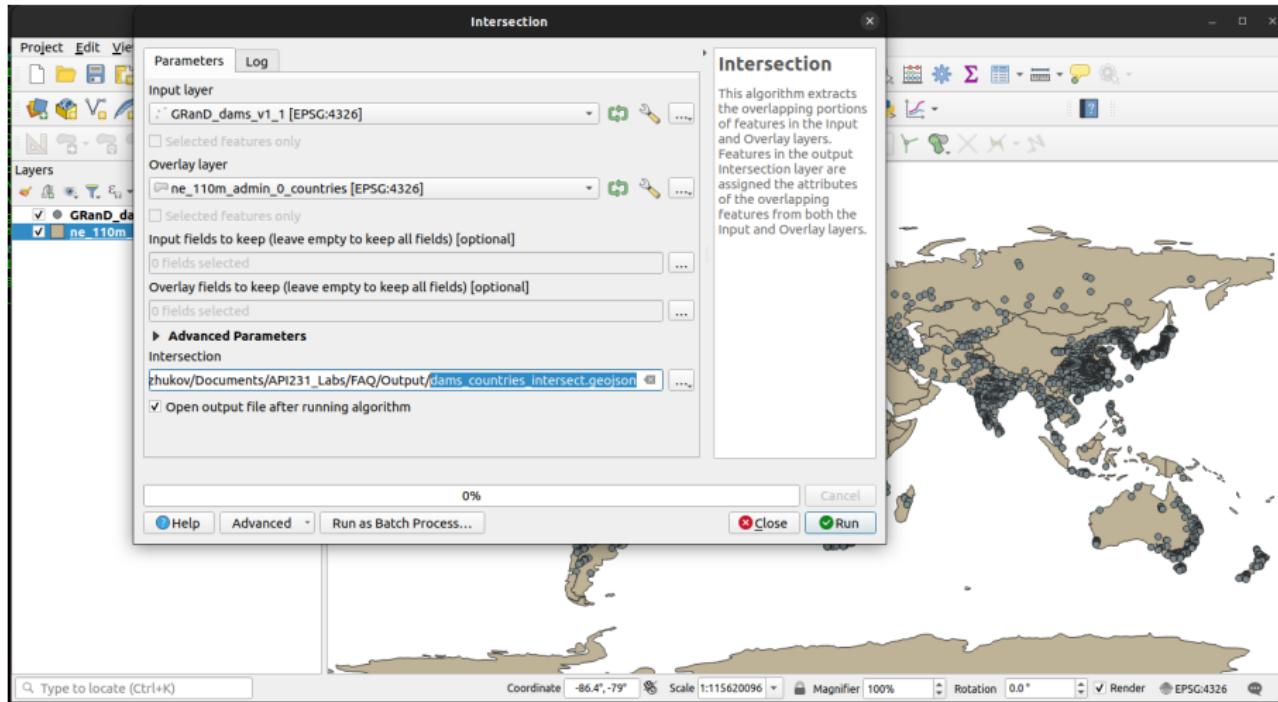
There are 2 ways to do this in QGIS: Intersection or Join attributes by location. Let's demonstrate here with data we've used before on *dams* (points) and *country borders* (polygons defining areas/categories).



The Intersection tool (Vector → Geoprocessing tools) will assign the attributes of the polygon that intersects with each point, while dropping points that fall outside the polygons.



Select the point layer as the Input layer and the polygons as Overlay layer, and adjust the overlay fields to keep/drop as needed.



If we compare the attribute tables of the intersection (top) vs. the original (bottom), we see that the intersection contains multiple additional columns from the polygon layer (e.g. ADMIN, ADM0_A3, etc.), while the original ends with LAT_DD.

dams_countries_intersect — Features Total: 6832, Filtered: 6832, Selected: 0

	LONG_DD	LAT_DD	featurecla	scalerank	LABELRANK	SOVEREIGNT	SOV_A3	ADM0_DIF	LEVEL	TYPE	TLC	ADMIN	ADM0_A3	GEO_U
1	-153.027083	57.65125	Admin-0 co...		1	2 United Stat...	US1		1	2 Country	1	United Stat...	USA	
2	-135.362917	63.774583	Admin-0 co...		1	2 Canada	CAN		0	2 Sovereign c...	1	Canada	CAN	
3	-133.72875	58.170417	Admin-0 co...		1	2 United Stat...	US1		1	2 Country	1	United Stat...	USA	
4	-122.199583	56.020417	Admin-0 co...		1	2 Canada	CAN		0	2 Sovereign c...	1	Canada	CAN	
5	-121.987083	55.99125	Admin-0 co...		1	2 Canada	CAN		0	2 Sovereign c...	1	Canada	CAN	
6	-131.342917	55.615417	Admin-0 co...		1	2 United Stat...	US1		1	2 Country	1	United Stat...	USA	
7	-129.857917	55.445417	Admin-0 co...		1	2 Canada	CAN		0	2 Sovereign c...	1	Canada	CAN	
8	-131.524583	55.379583	Admin-0 co...		1	2 United Stat...	US1		1	2 Country	1	United Stat...	USA	
9	-126.224583	54.807083	Admin-0 co...		1	2 Canada	CAN		0	2 Sovereign c...	1	Canada	CAN	

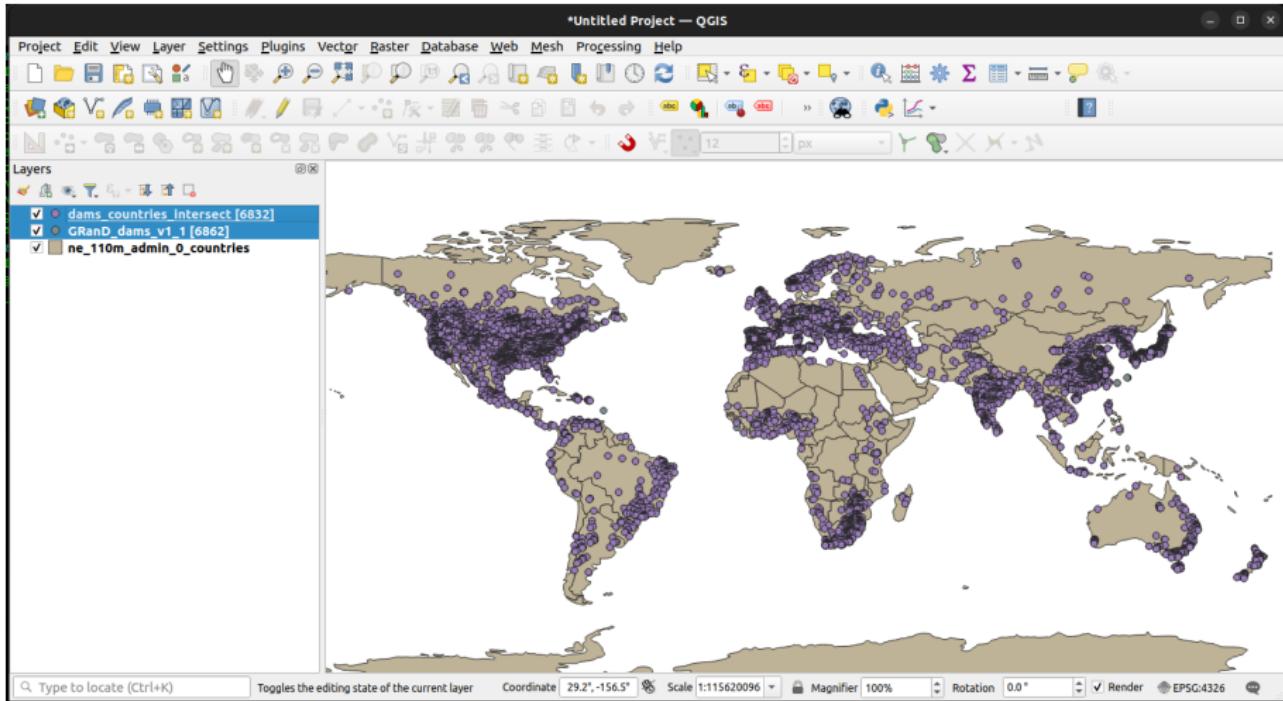
Show All Features ↴

GRanD_dams_v1_1 — Features Total: 6862, Filtered: 6862, Selected: 0

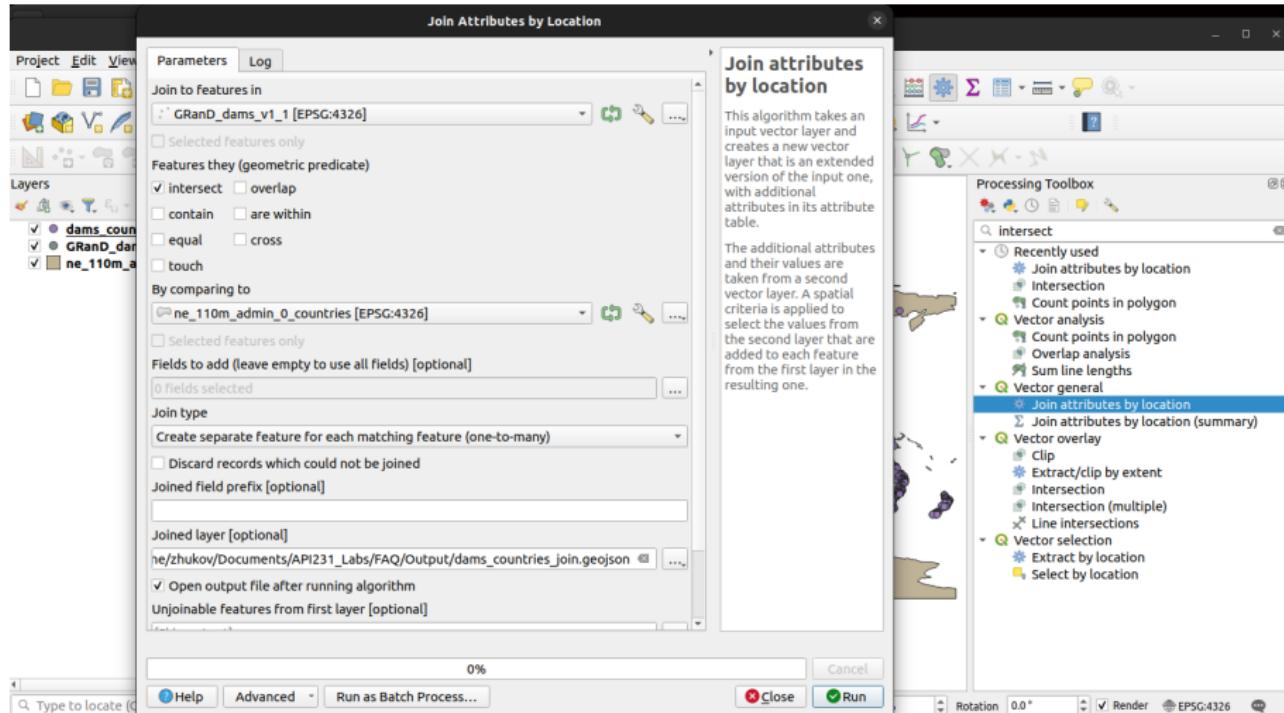
	USE_FISH	USE_PCON	USE_LIVE	USE_OTHR	MAIN_USE	LAKE_CTRL	MULTI_DAMS	TIMELINE	COMMENTS	URL	QUALITY	EDITOR	LONG_DD	LAT_DD
1	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	http://ww...	3: Fair	UNH	-153.027083	57.651250
2	JLL	NULL	NULL	NULL	Hydroelect...	Yes	NULL	NULL	This dam is ...	http://ww...	1: Verified	McGill	-135.362917	63.774583
3	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	http://ww...	2: Good	UNH	-135.199583	57.066250
4	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	http://ww...	2: Good	UNH	-135.110417	56.990417
5	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	http://ww...	4: Poor	UNH	-133.728750	58.170417
6	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	Polygon to...	NULL	3: Fair	McGill	-122.199583	56.020417
7	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	NULL	3: Fair	McGill	-121.987083	55.991250
8	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	http://ww...	3: Fair	UNH	-131.342917	55.615417
9	JLL	NULL	NULL	NULL	Hydroelect...	NULL	NULL	NULL	NULL	NULL	3: Fair	McGill	-129.857917	55.445417

Show All Features ↴

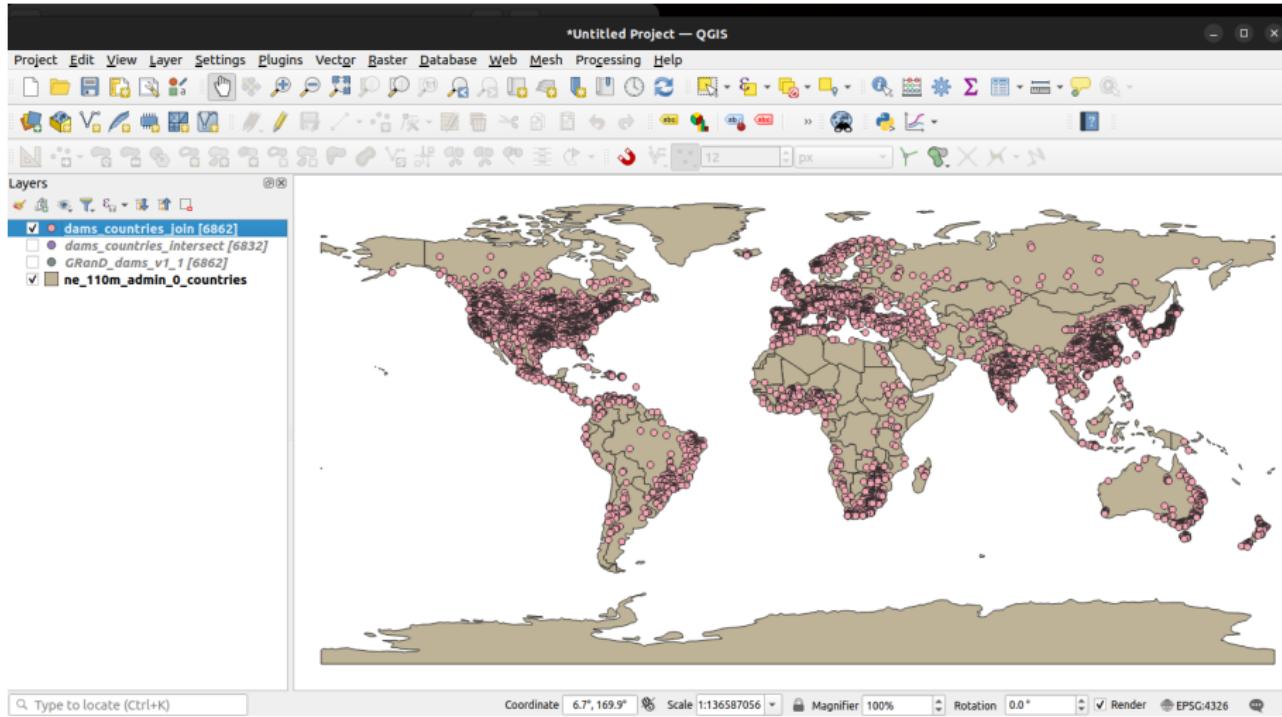
However, the feature count in the layer menu tells us that the intersect layer contains 6832 features (points), but the original dams layer contained 6862. So we lost 30 dams that fell outside of all national borders. What if we want to keep them?



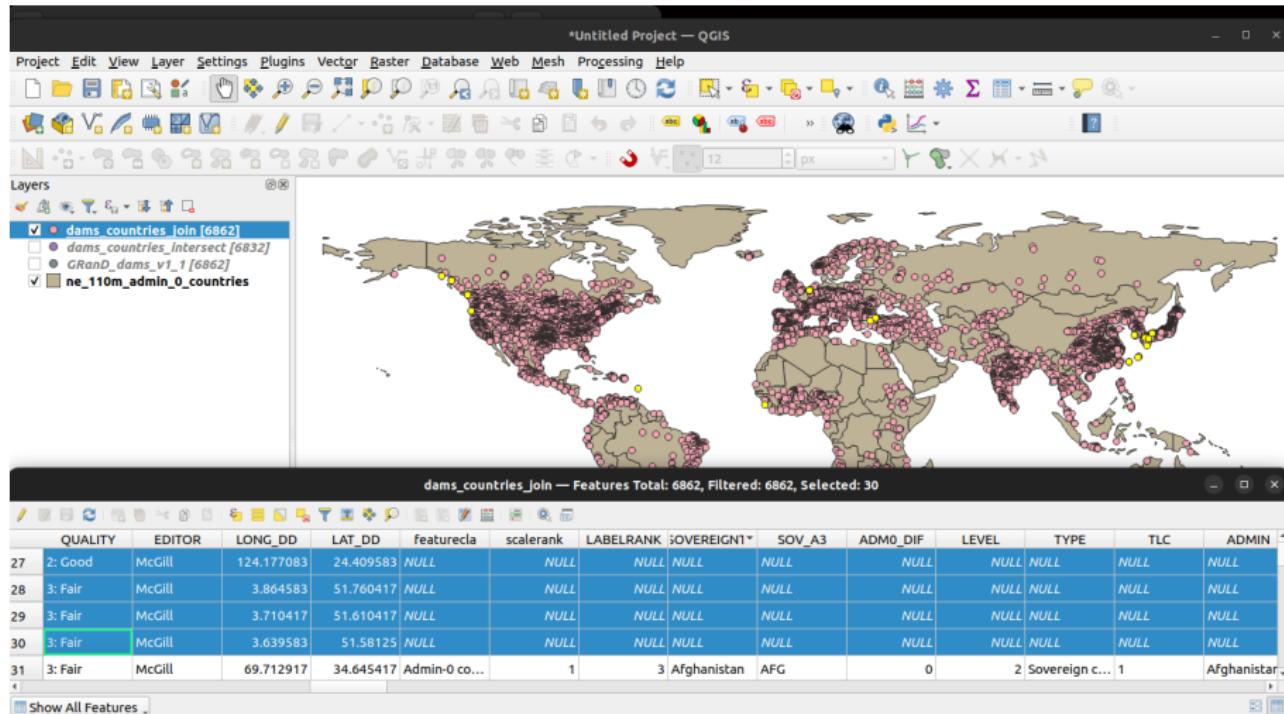
The other option is to use the **Join Attributes by Location** tool (Processing Toolbox → Vector general). It's the same idea, but with more options (like whether to "Discard records which could not be joined")



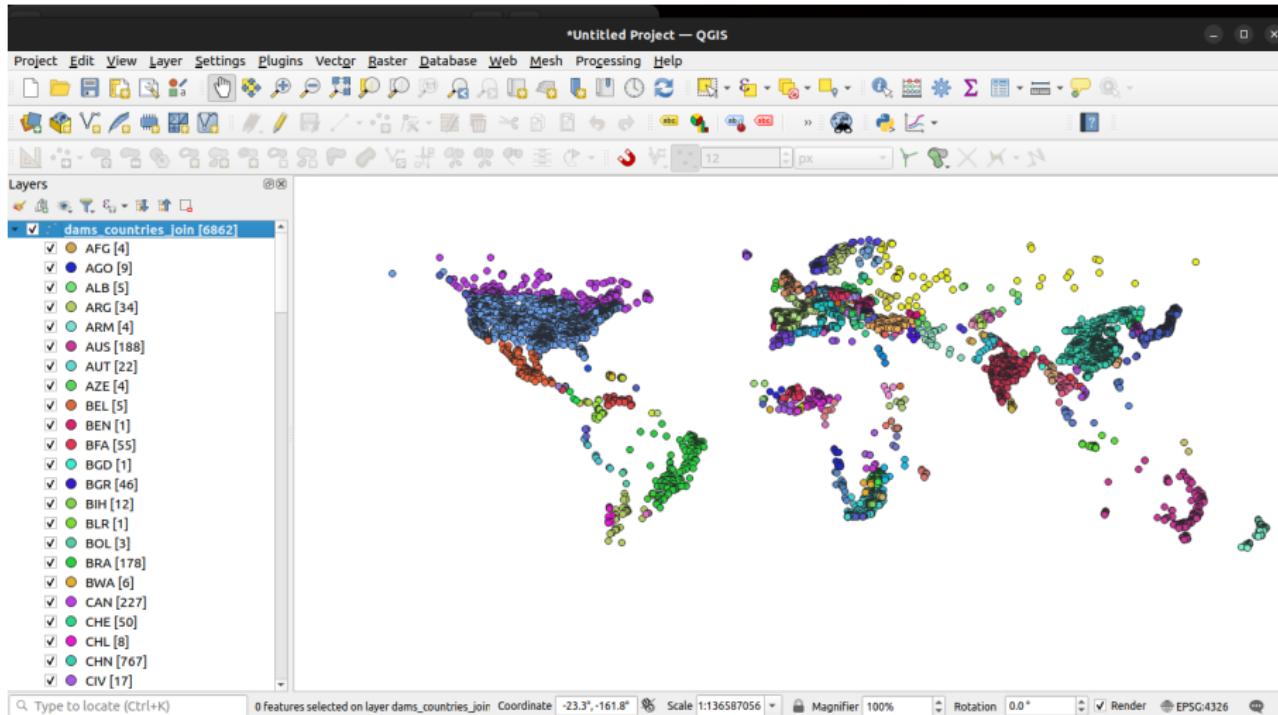
The feature count for the joined layer is the same as for the original dams layer (6862), as long as “Discard records which could not be joined” is unchecked.



The points that intersect with no polygons are given NULL values for the joined fields. We can select them, and see that most of these are in coastal waters or on islands in/near international waters.



You can plot the joined attributes to be sure that everything worked out as expected



Visualizing comparisons between maps

“What is the best recommended way to show comparison between maps? ... what mapping techniques do you recommend for displaying two different datasets on one map? (e.g. Climate and violence, or voting preferences and income per capita).”

If one variable is *continuous* (e.g. income) and the other is *categorical* (e.g. yes/no), you can use a gradient for the continuous variable, and shading lines for the categorical one (QGIS: Single Symbol / Hashed; R: `plot(..., density=15, angle=30)`). You'll need to duplicate the layer to display > 1 variable at a time

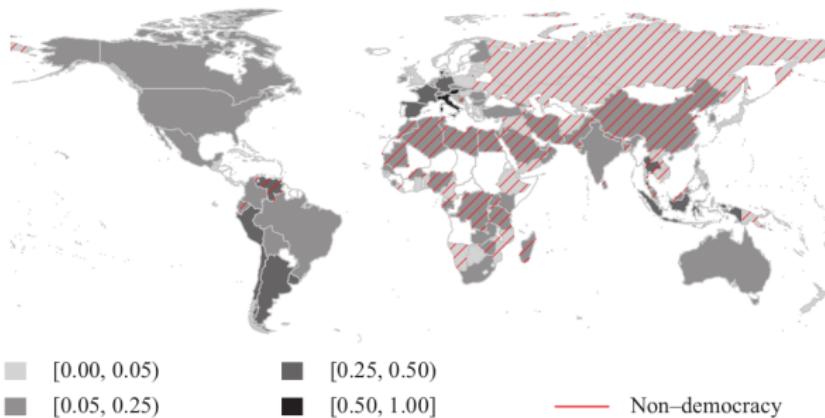
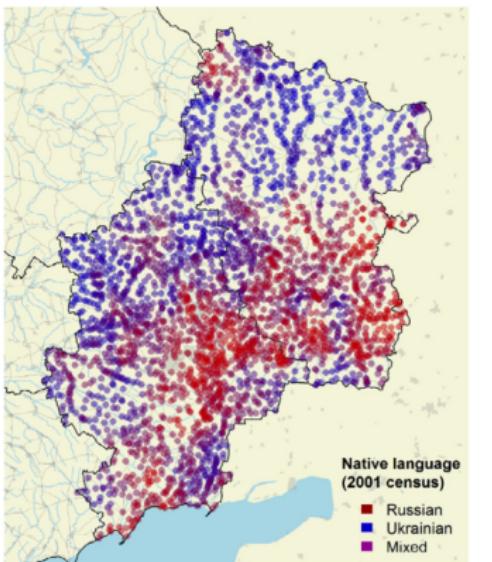


Figure 2. Libya news coverage data

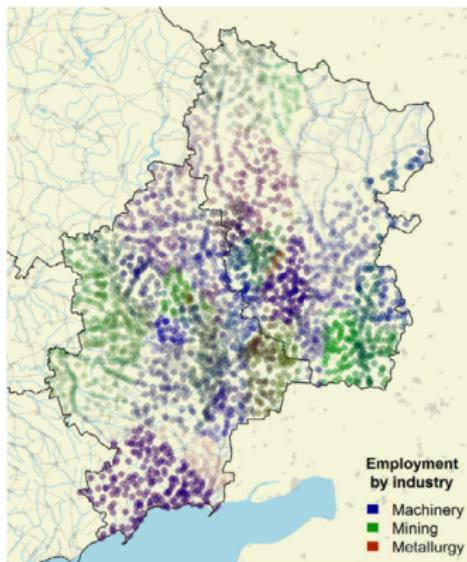
Frequency of newspaper reports on Libyan crisis. Shadings correspond to proportion of newspaper-days with at least one article published.

Figure 1: Example from Baum and Zhukov (2015)

If *both* variables are *continuous*, it is better to create two maps side-by-side.



(a) Russian language



(b) Local employment

Figure 2: Example from Zhukov (2016)

When you do this, make sure the map extent is the same for both maps, and keep everything identical except for the variables you want to compare.

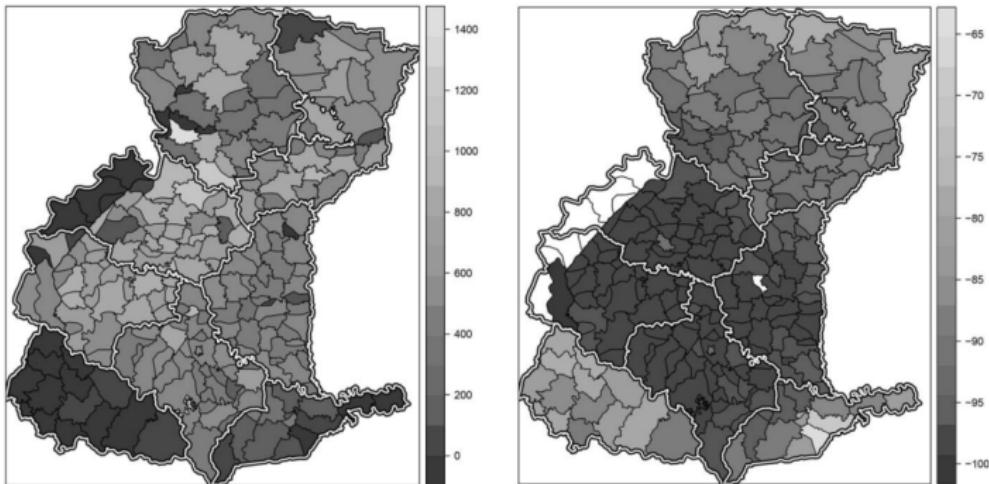


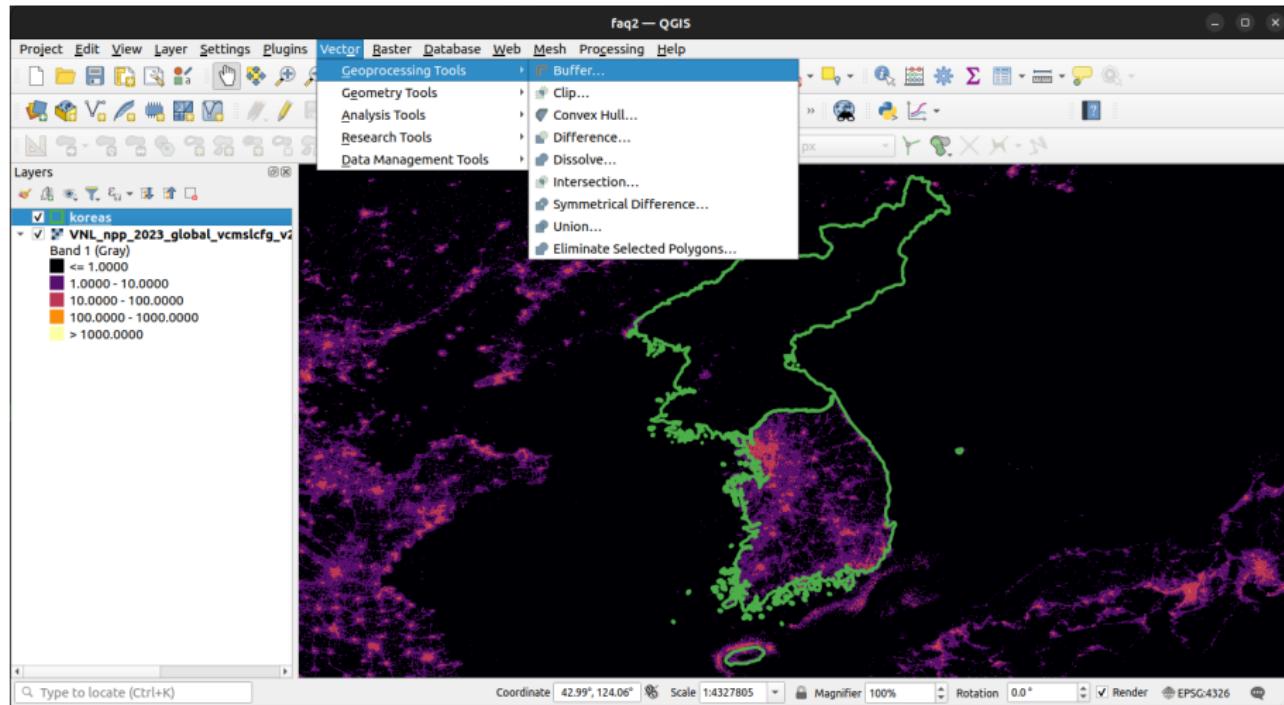
Figure 1. Historical violence and contemporary voting in western Ukraine. The figure on the left shows the counts of deported individuals. The right panel shows the pro-Russian vote margin in the 2014 parliamentary elections. The westernmost rayons in white have no election data because the USSR returned them to Poland in 1945. Historical boundaries of oblasts appear in white. Please refer to appendix 2 for residualized maps that account for systematic regional differences.

Figure 3: Example from Rozenas et al (2017)

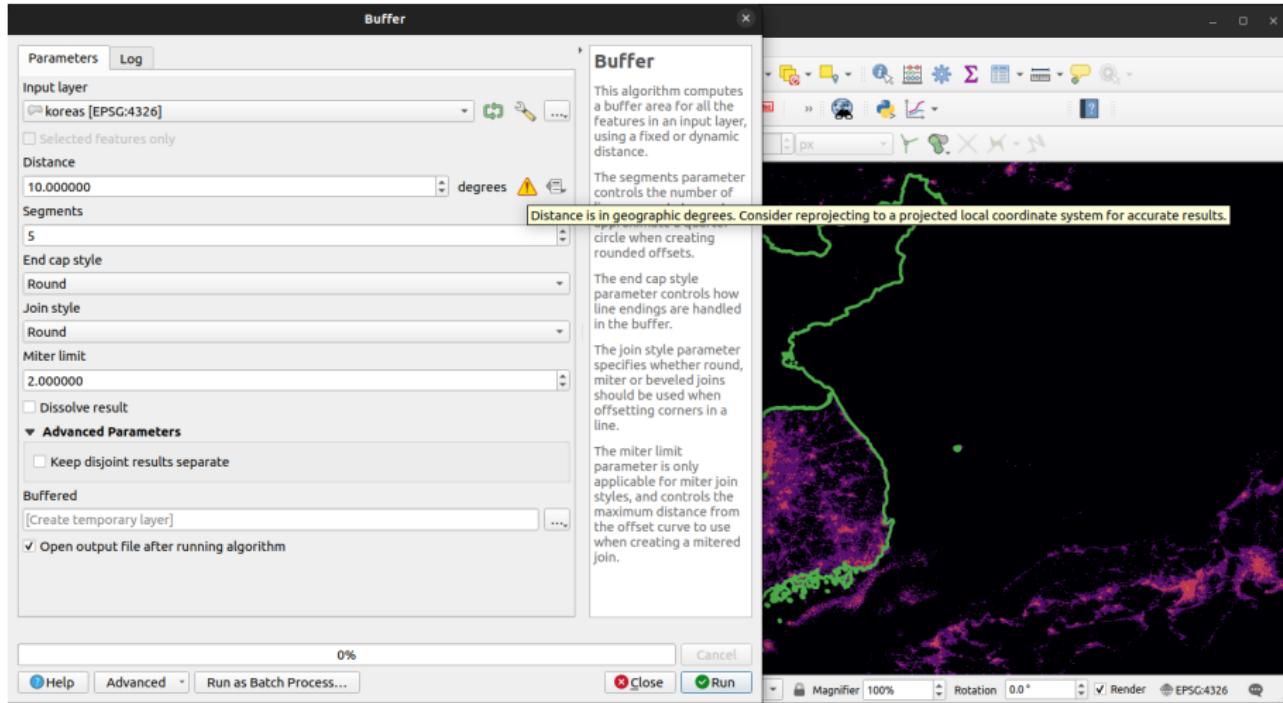
Buffers

“Is there a way to create a safety margin when using raster extraction by mask?
(Say, I want an additional 10km outside the boundary of my vector layer to be cropped as well)”

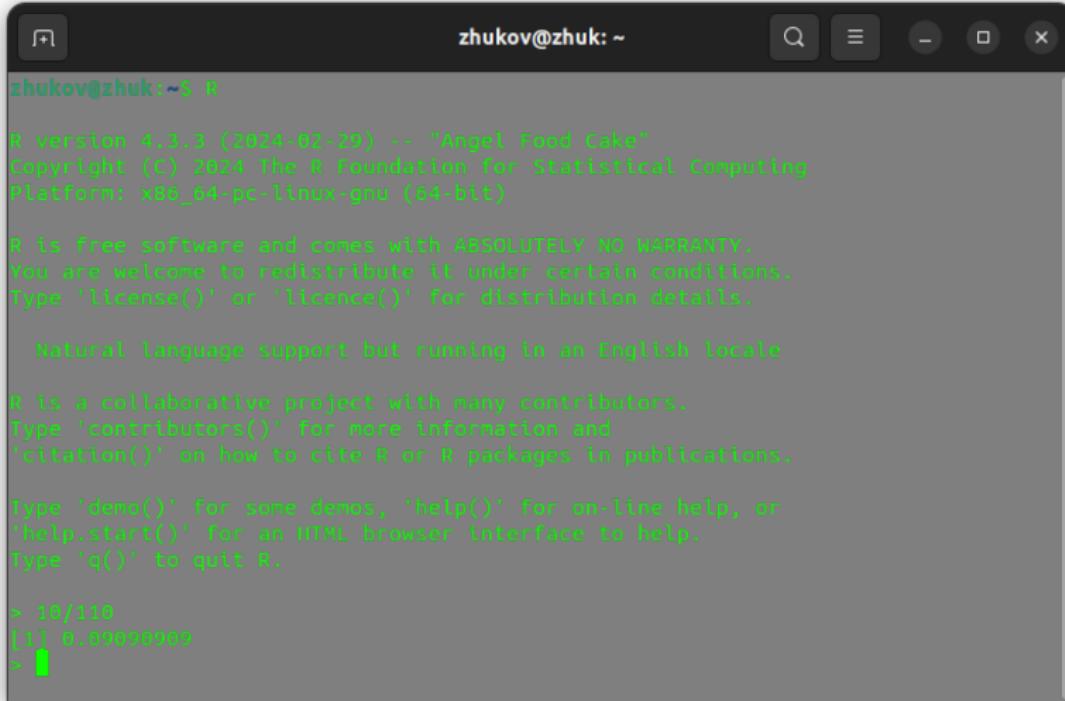
There is no way to do this within Zonal statistics directly, but we can use the Buffer tool to pre-process the polygon layer. Let's demonstrate with data on *luminosity* (raster) and *country borders* (polygons).



If the input layer (`koreas`) is unprojected, the Buffer tool will ask for the distance in degrees. You can either change the CRS or do some back-of-the-envelope math.



1 degree \approx 100 km at the equator. If we want a 10km buffer, that's roughly $10/110 = .091$ degrees.



The screenshot shows a terminal window titled "zhukov@zhuk: ~". The window contains the following text:

```
R version 4.3.3 (2024-02-29) -- "Angel Food Cake"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

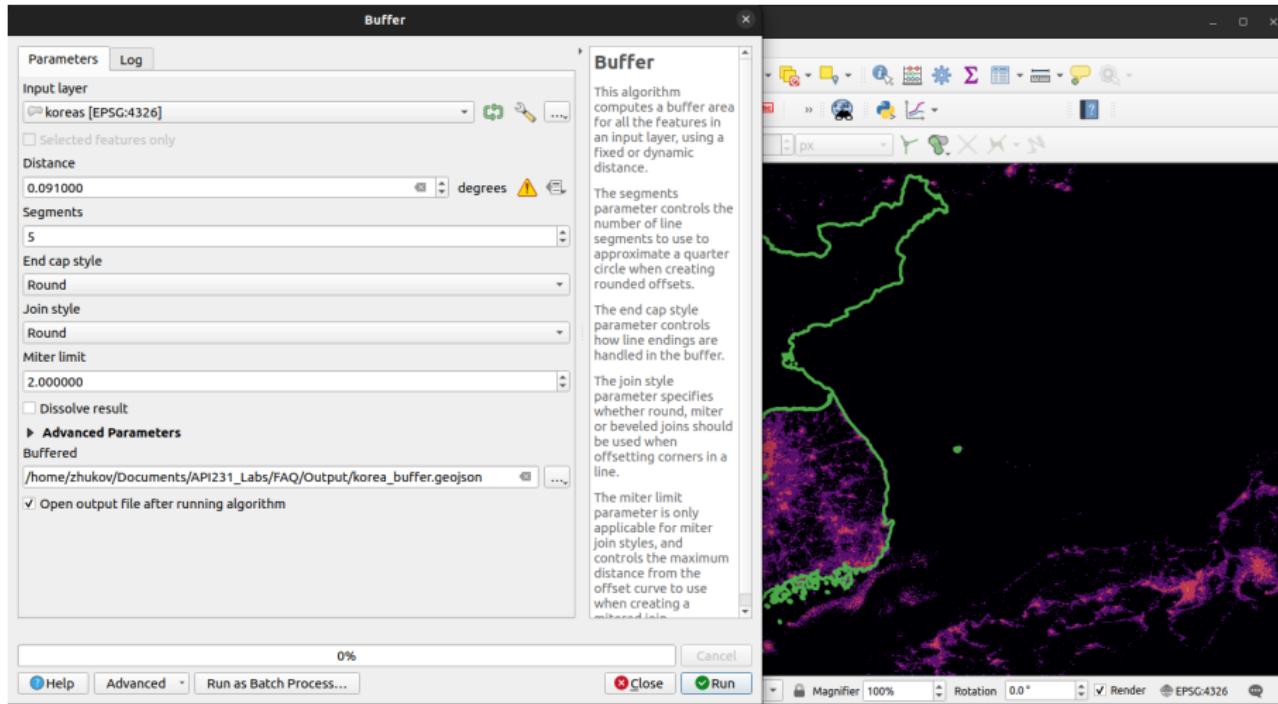
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

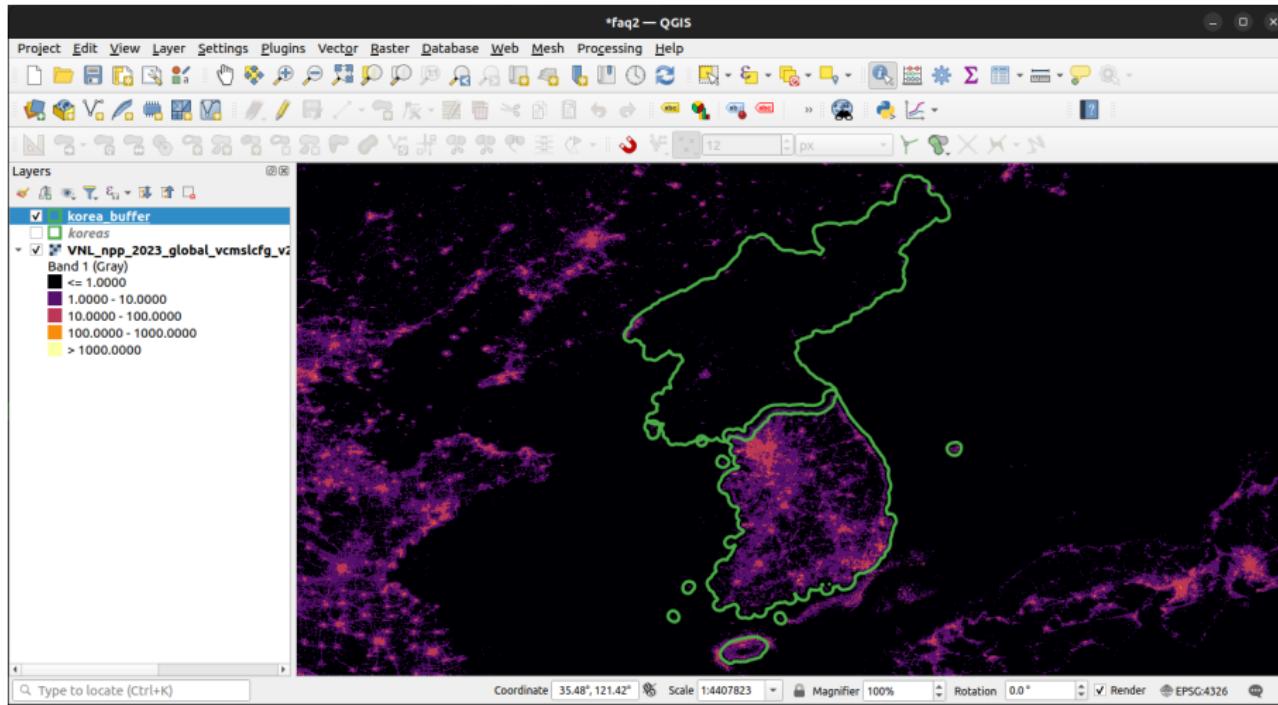
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 10/110
[1] 0.09090909
> █
```

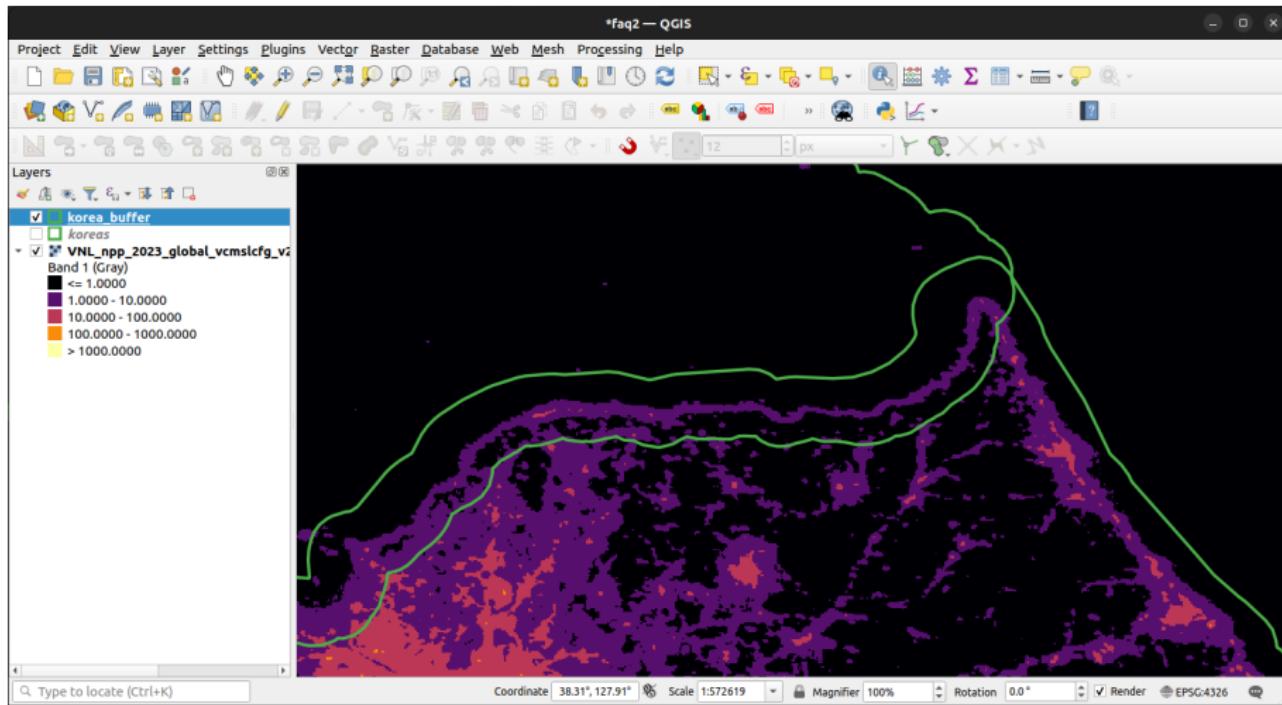
Enter the converted distance in Distance and run the buffer tool.



The buffered polygon should look similar, but “puffier”



Note that the buffers will *overlap* in neighboring polygons. So, North Korea will include 10km of South Korea and vice versa.



There are tutorials and YouTube videos online on how to remove the overlap, but it's too complex to cover here.

The screenshot shows a Stack Exchange question page for the 'Geographic Information Systems' category. The question title is "Buffer neighbouring polygons without overlap using QGIS". It was asked 8 years, 3 months ago and modified 2 years ago, with 5k views. The question asks if it's possible to create a buffer around neighbouring polygons so that they do not overlap, preferably in QGIS. Below the question, there's an image showing two sets of overlapping polygons. The left set shows three green polygons overlapping within a larger orange polygon. The right set shows the same three green polygons, but each is buffered individually, resulting in three separate, non-overlapping orange buffers. A sidebar on the right includes links to "The Overflow Blog" and "Featured on Meta".

Buffer neighbouring polygons without overlap using QGIS

Asked 8 years, 3 months ago Modified 2 years ago Viewed 5k times

Is it possible to create a buffer around neighbouring polygons, so that the new polygons do not overlap?

27 Preferably in QGIS but any other tool will do.

So instead of the first result I would like to obtain something similar to the second:

The Overflow Blog

- Diverting more backdoor disasters
- How to succeed as a data engineer without the burnout

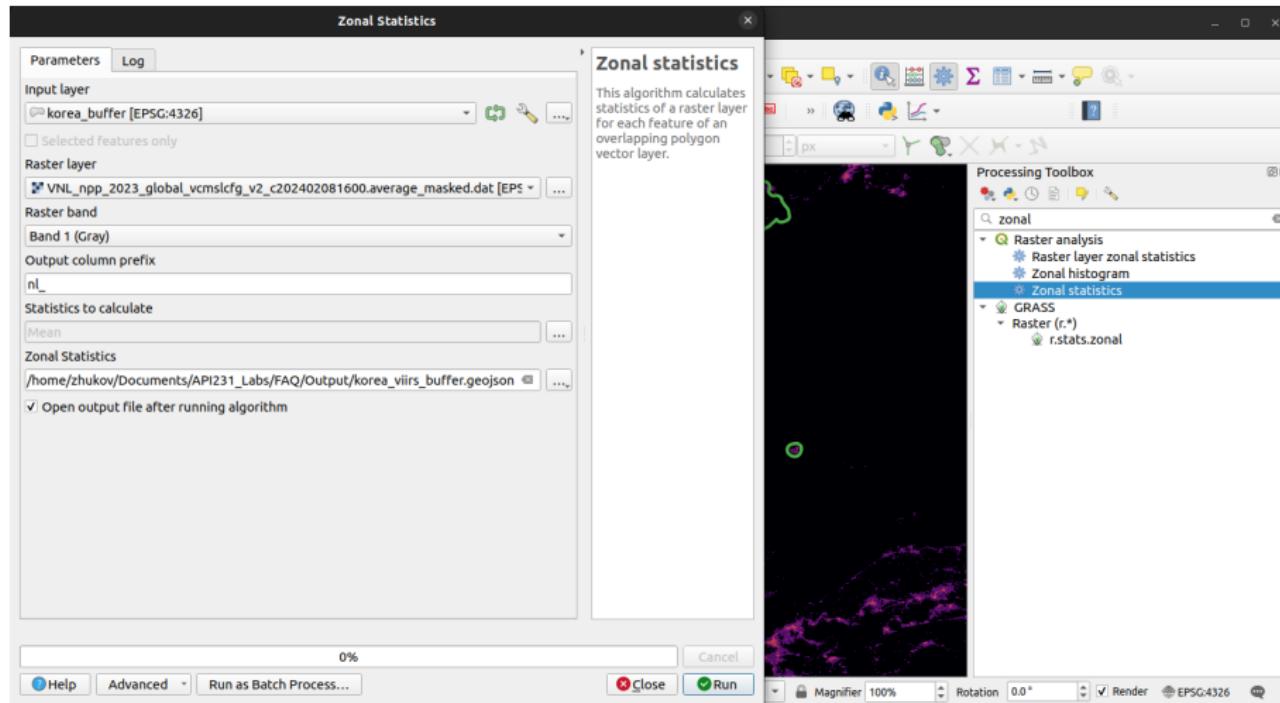
Featured on Meta

- New Focus Styles & Updated Styling for Button Groups
- Upcoming initiatives on Stack Overflow and across the Stack Exchange network

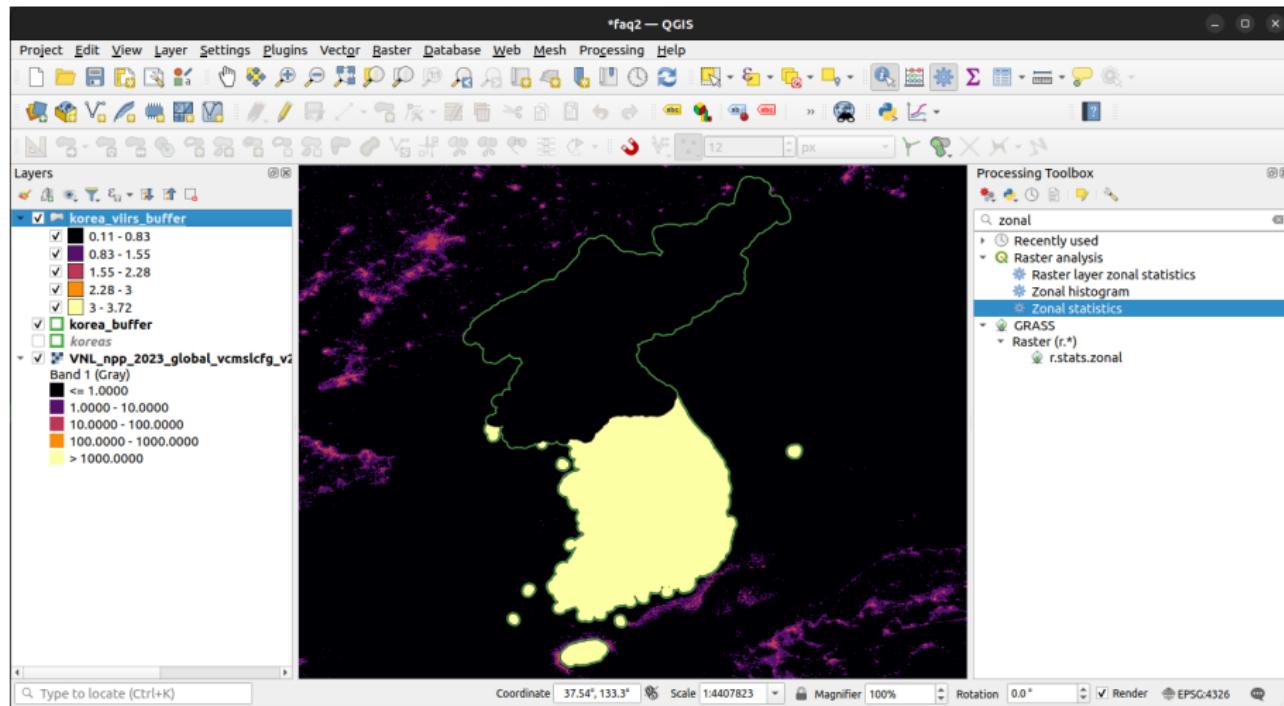
Linked

Creating Thiessen (Voronoi) polygons using lines (rather than points) as the input features?

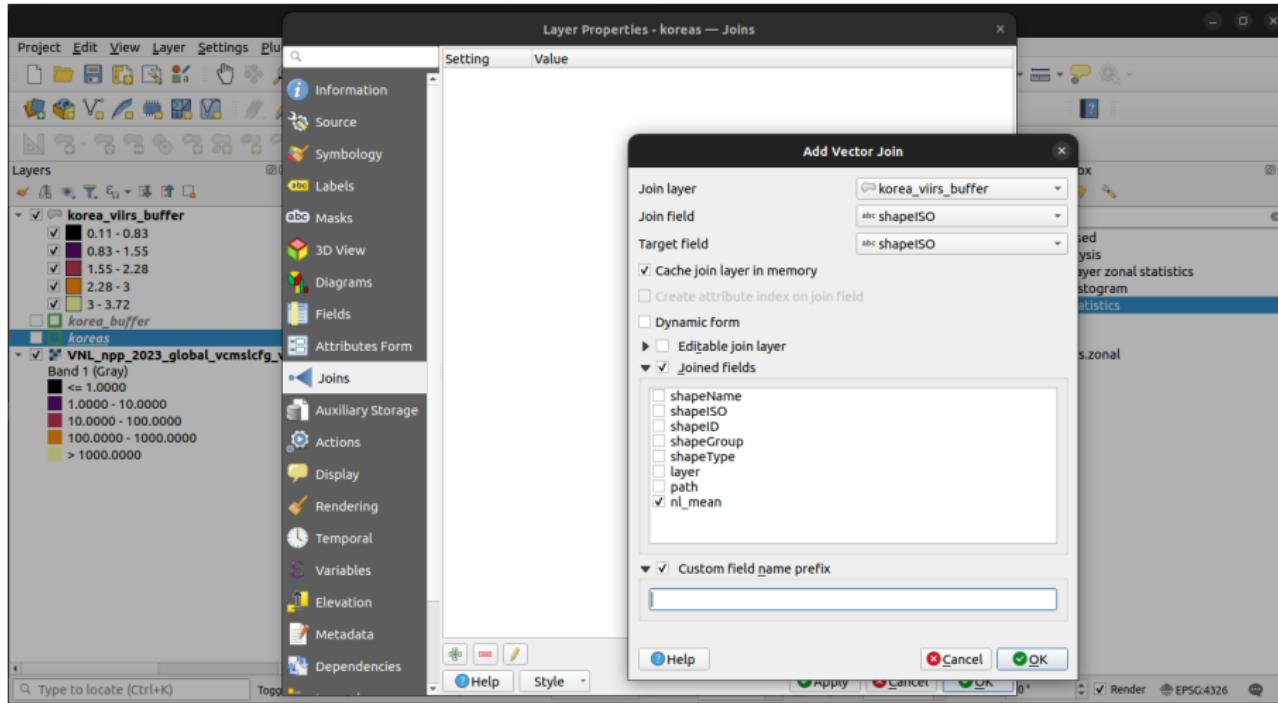
You can now implement Zonal statistics with the buffered polygons as the Input layer.



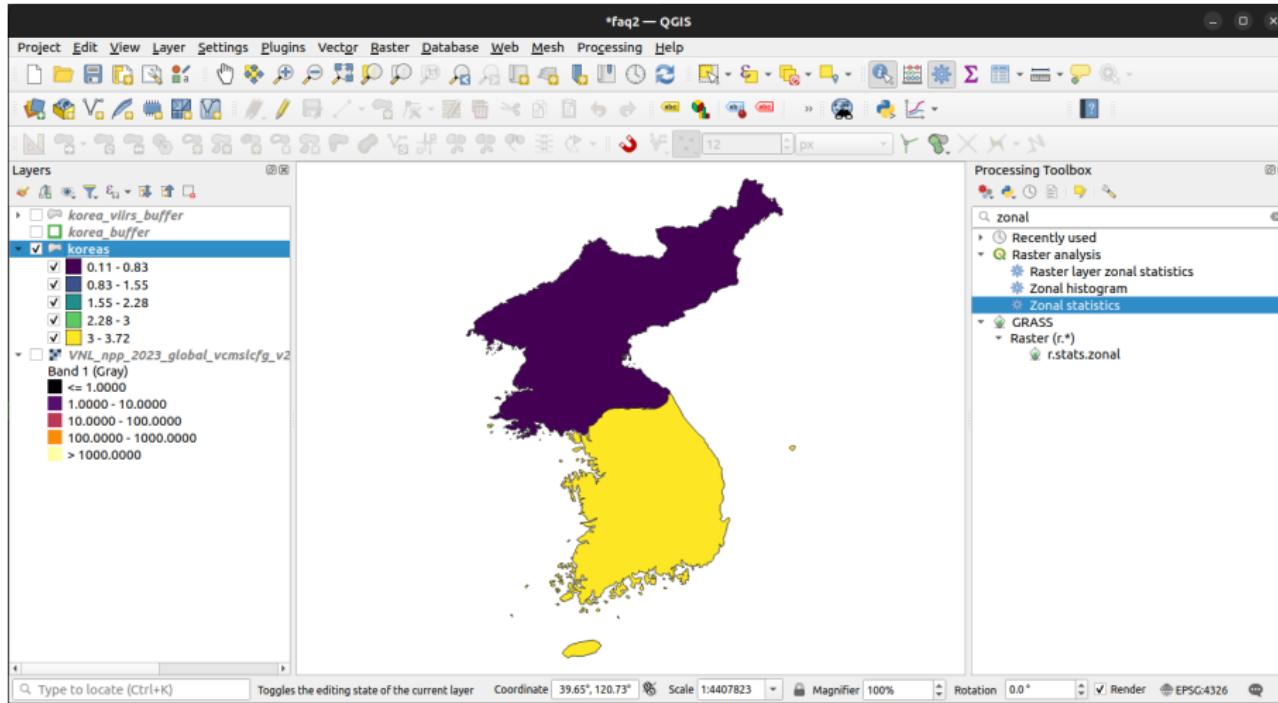
You can plot the mean luminosity, and confirm that South Korea is brighter than the North. But you may also want to merge the results back to the original, non-buffered polygons



You can do this by adding a Vector join in layer properties, here adding the nl_mean variable from the buffered polygons to the original polygons.



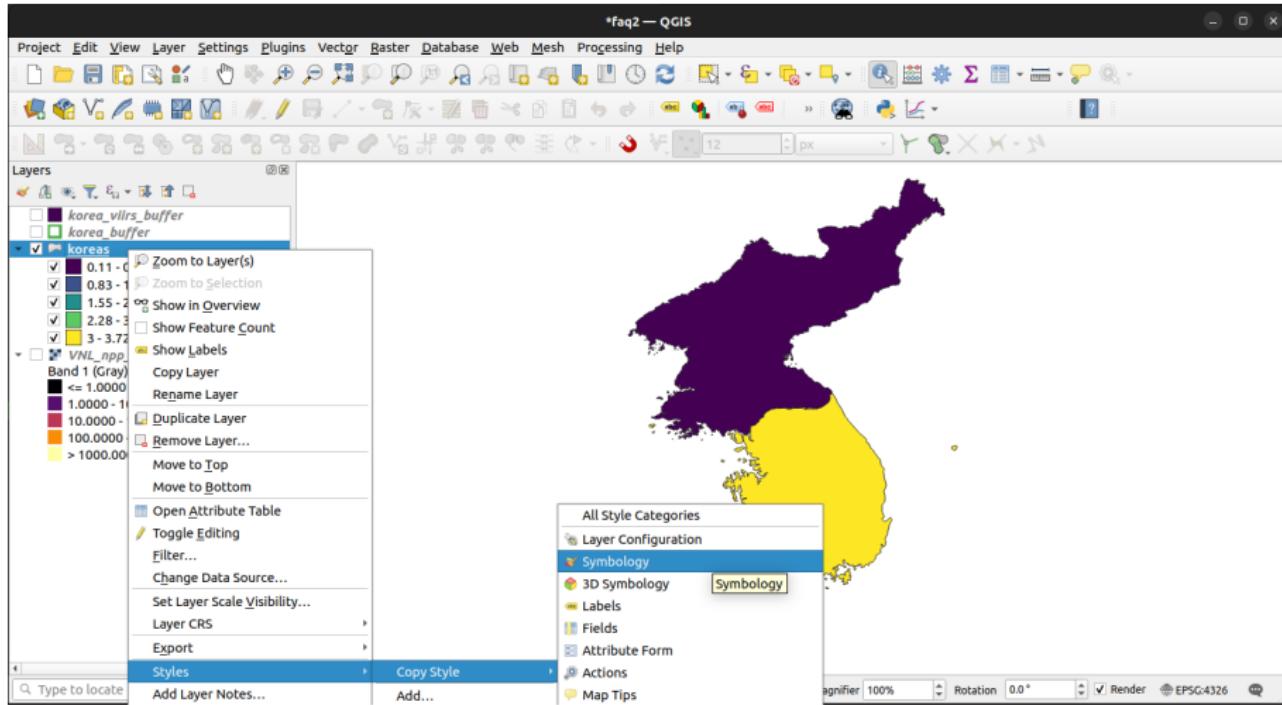
This way, you get to keep the original polygon geometries, while using buffered geometries to calculate zonal statistics.



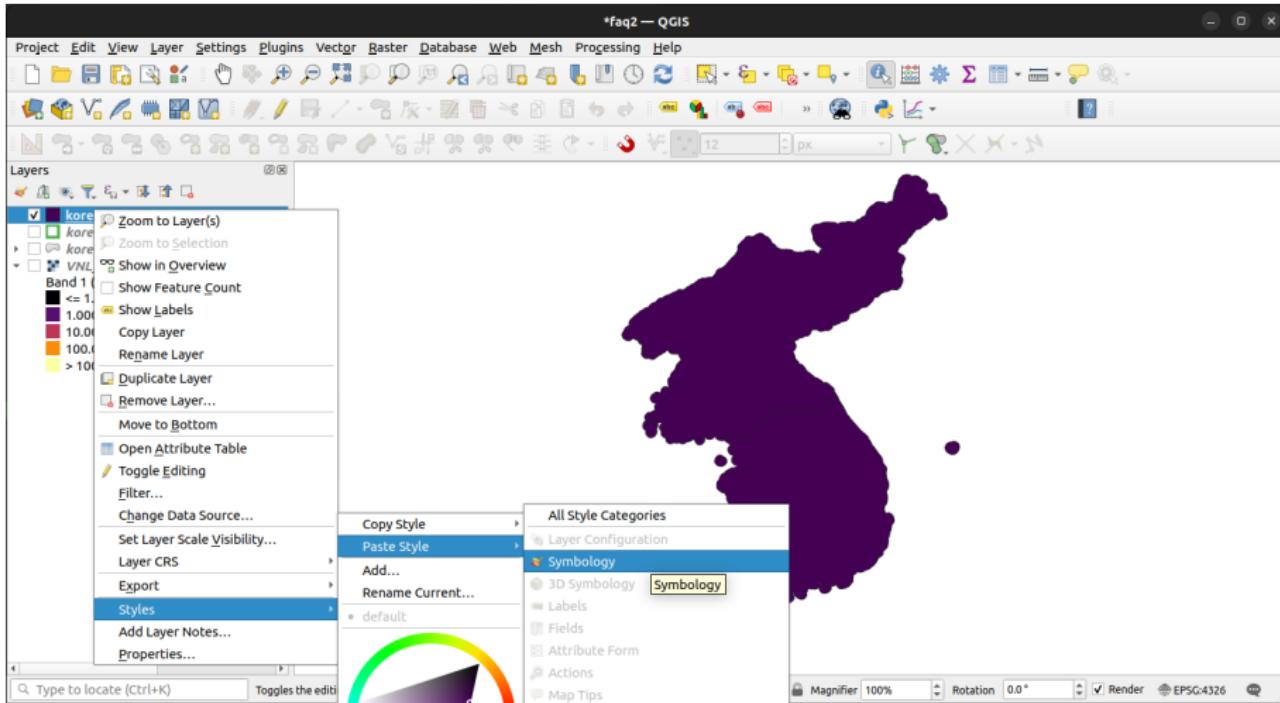
Re-using the same color symbology

“How do I duplicate a color scale across different layers of my project?”

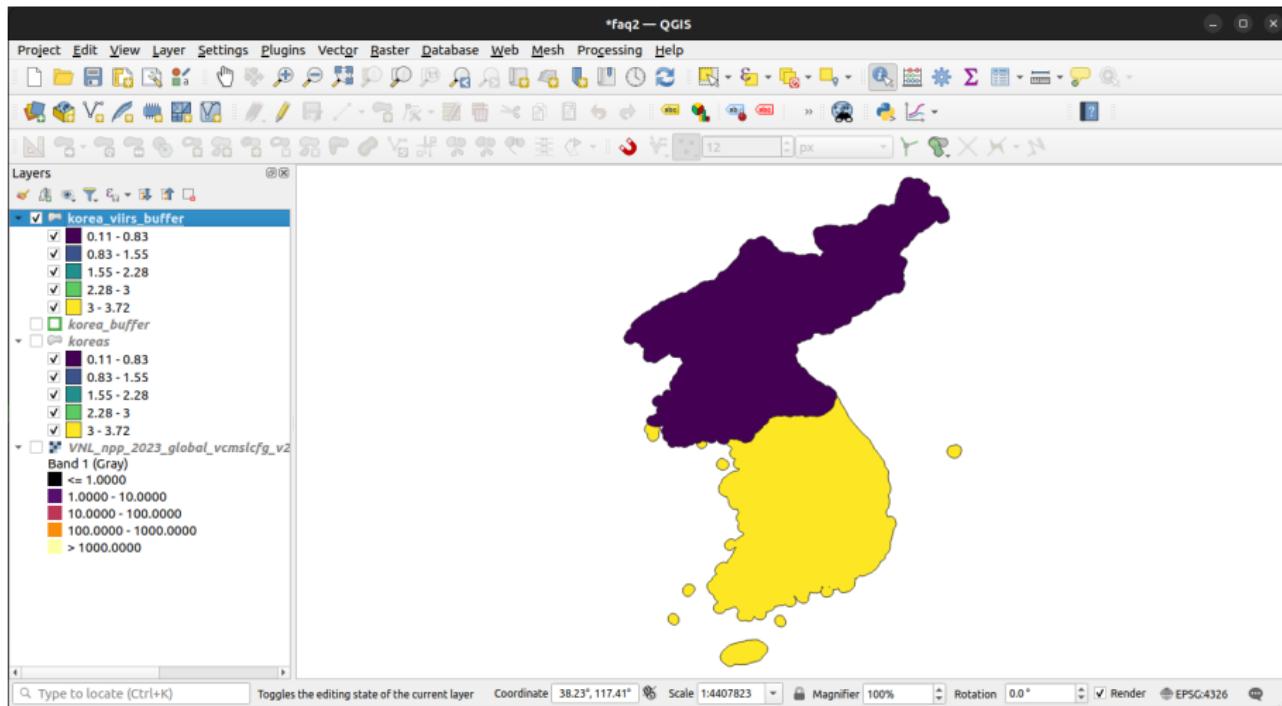
In QGIS, right-click on the layer whose color symbology you want to duplicate, and select Styles → Copy style → Symbology



Now right-click on the layer whose color symbology you want to replace, and select **Styles → Paste style → Symbology**



The color scheme and break points should now be replicated in the second layer.
Note that you will still need to re-classify the colors in Properties if the numerical distribution is different in the second layer.



Regression analysis

"I am intrigued by regression analysis and its application. As a student with limited experience in statistics or regression analysis, I wonder if it is feasible for someone like myself to undertake a basic level of analysis."

Flashback: we used regression analysis in Walk Through 1 (Islamic State):

$$\begin{aligned}\text{violence}_i = & \beta_1 \text{road density}_i + \beta_2 \text{population}_i + \beta_3 \text{cropland}_i \\ & + \beta_4 \text{dams}_i + \beta_5 \text{Sunni presence}_i + \epsilon_i\end{aligned}$$

where

- violence_i was the observed number of ISIS attacks in district i
- $\text{road density}_i, \dots, \text{Sunni presence}_i$ were explanatory variables
- ϵ_i were errors (residuals)
- β were coefficient estimates corresponding to each Hypothesis

Hypothesis	Expectation	Observation
1. Power projection	$\beta_1 < 0$?
2. Demographics	$\beta_2 > 0$?
3. Political economy	$\beta_3 < 0$?
4. Key infrastructure	$\beta_4 > 0$?
5. Sectarian divisions	$\beta_5 > 0$?

Several popular types of (basic) regression models

Model	Type of dependent variable	R command
1. Linear regression (OLS)	continuous (0.47, -1.97, -0.29)	<code>lm()</code>
2. Logistic regression (logit)	binary (0, 1)	<code>glm(..., family="binomial")</code>
3. Quasi-Poisson	counts (0, 1, 2, 3, ...)	<code>glm(..., family="quasipoisson")</code>

Online tutorials (partial list)

1. Free: Princeton library guides
 - a) Linear: libguides.princeton.edu/R-linear_regression
 - b) Logit: libguides.princeton.edu/R-logit
2. Paid/subscription: DataCamp
 - a) Linear: datacamp.com/tutorial/linear-regression-R
 - b) Logit: datacamp.com/tutorial/logistic-regression-R