CODECHECK certificate 2022-001

https://doi.org/10.5281/zenodo.6040066

Item	Value
Title	Geographically Weighted Regressions for prioritizing educational planning, policies, and interventions
Authors	Germán Vargas Mesa, Amélie A. Gagnon
Reference	IIEP technical note (2021) http://www.iiep.unesco.org/en/publication/geographically-weighted-regressions-prioritizing-educational-planning-policies-and
Codechecker	Stephen J. Eglen
Date of check	2022-01-19 10:00:00
Summary	R code for this project was reproducible.
Repository	https://github.com/codecheckers/GWR-in-educational-planning

Table 1: CODECHECK summary

Output	Comment	Size (b)
codecheck/results/fig4.pdf	manuscript Figure 4	11040
codecheck/results/fig5.pdf	manuscript Figure 5	2044326
codecheck/results/fig6.pdf	manuscript Figure 6	1789437
codecheck/results/fig7.pdf	manuscript Figure 7	2735502
codecheck/results/fig8.pdf	manuscript Figure 8	33975
codecheck/results/table2.csv	Data underlying manuscript Table 2	2979
codecheck/results/table3.csv	Data underlying manuscript Table 3	1445
codecheck/results/code.Rout	Text output from R code (not in manuscript)	54601
codecheck/results/Rplots.pdf	Graphical output from R code (not in manuscript)	75407993

Table 2: Summary of output files generated

Summary

The code could re-run successfully.

Key challenges: (1) dataset required is large and access is provided on request rather than being freely available on the internet. (2) significant number of R packages to install – but all are available and so just requires some time to set up, together with corresponding unix binaries. (3) To visualise the results, the user needs to be familiar with the QGIS application.

CODECHECKER notes

The GitHub repo was https://github.com/codecheckers/GWR-in-educational-planning

Installation prerequisites

A file `codecheck/installs.R` was created to do the local installations. Some of the R packages required extra linux packages to be installed, notably 'gdal' and 'udunits' – see the R script for details. The installation required many R packages, taking about 20 minutes to install. This was non-trivial to setup, and perhaps in future could benefit from a Docker container.

Data

The data required for the project is not publicly available, but is available upon request (note procedure in github file). The file `Replication files.zip` is 867 Mb. This zip file was unpacked and stored in a separate directory to the github.

As the main R script was quite long, I created a symlink to the file

```
ln -s "Geographically weighted regressions for prioritizing educational planning, policies, and interve
```

Inside `code.R`, I set the variable

```
replication.folder = "/home/stephen/archive/proj/2022/gwr/Replication files"
```

Changes to the code.

Only minor changes to the code were required. I used `file.path()` rather than assuming the directory separator is \\ (as it is on Windows). I also added the followin line to the end of the R script so that it reports the packages used in the R at the end of code.Rout:

```
sessionInfo()
```

Running the code

Step 1: running the R code

To run the code:

```
R CMD BATCH code.R
```

I then used a script, `running.sh`, to copy key outputs across from the directory where code was run into the codecheck repo. The code took just over an hour to run. I have stored the `Rplots.pdf` and `code.Rout` output from the run into the results directory.

Step 2: running QGIS

The results from the R analysis are then visualised using QGIS. The author kindly made a video to show the steps required in generating the graphical output: https://youtu.be/9AGdQXMgsFo.

This video allowed me to reproduce the figures essentially in the same format. Note however the first few minutes discusses shapefiles that are not included in the repository and so the background of the images is different. Also, post-processing of the figures (e.g. the legend in the bottom left and the circle in the top-right) was not reproduced here. There are some minor differences observed however, for example in Figure 6, the statistically-signifanct regions in the middle of Colombia appear slightly different. Also, figure 7 seemed to be cropped at the top somehow, but the major points of the figure are present.

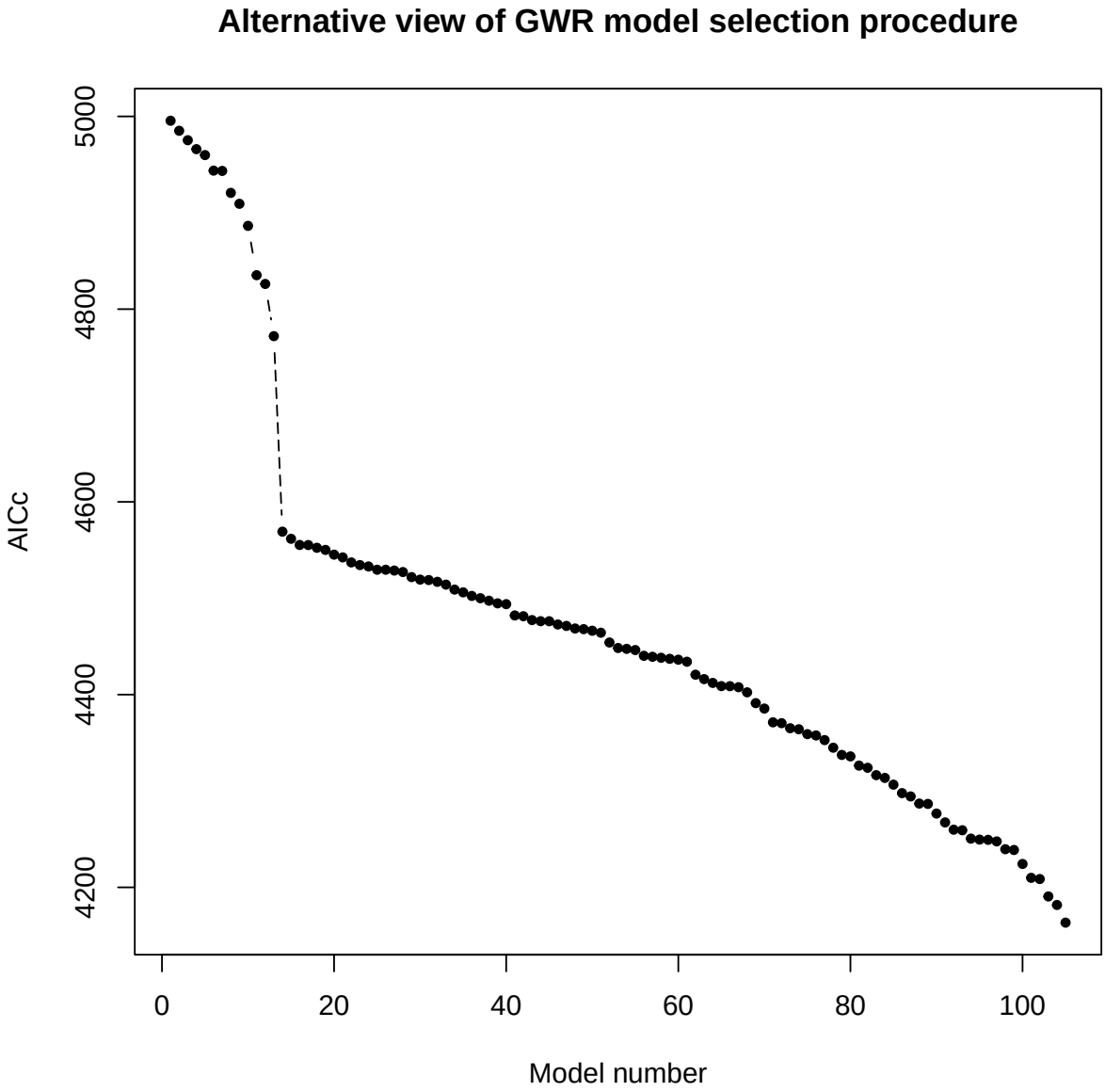

Figure C1: manuscript Figure 4

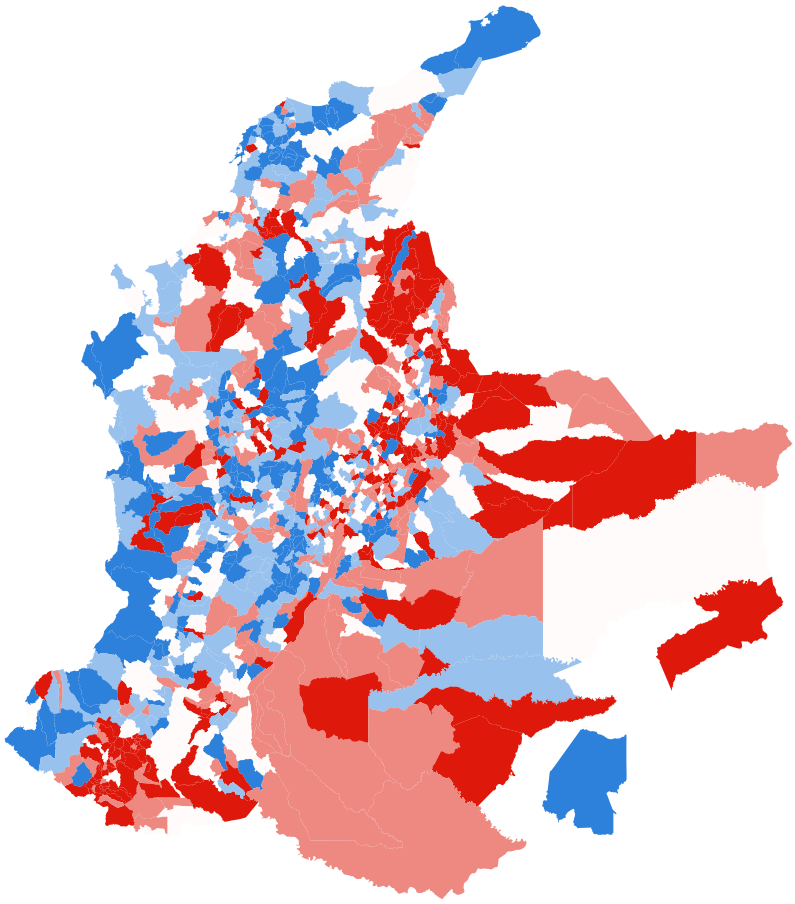

Figure C2: manuscript Figure 5

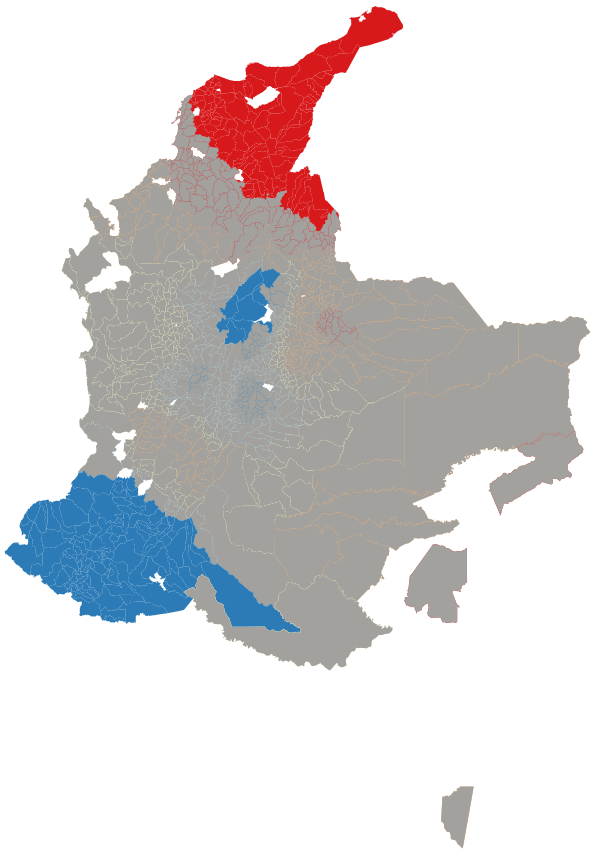

Figure C3: manuscript Figure 6

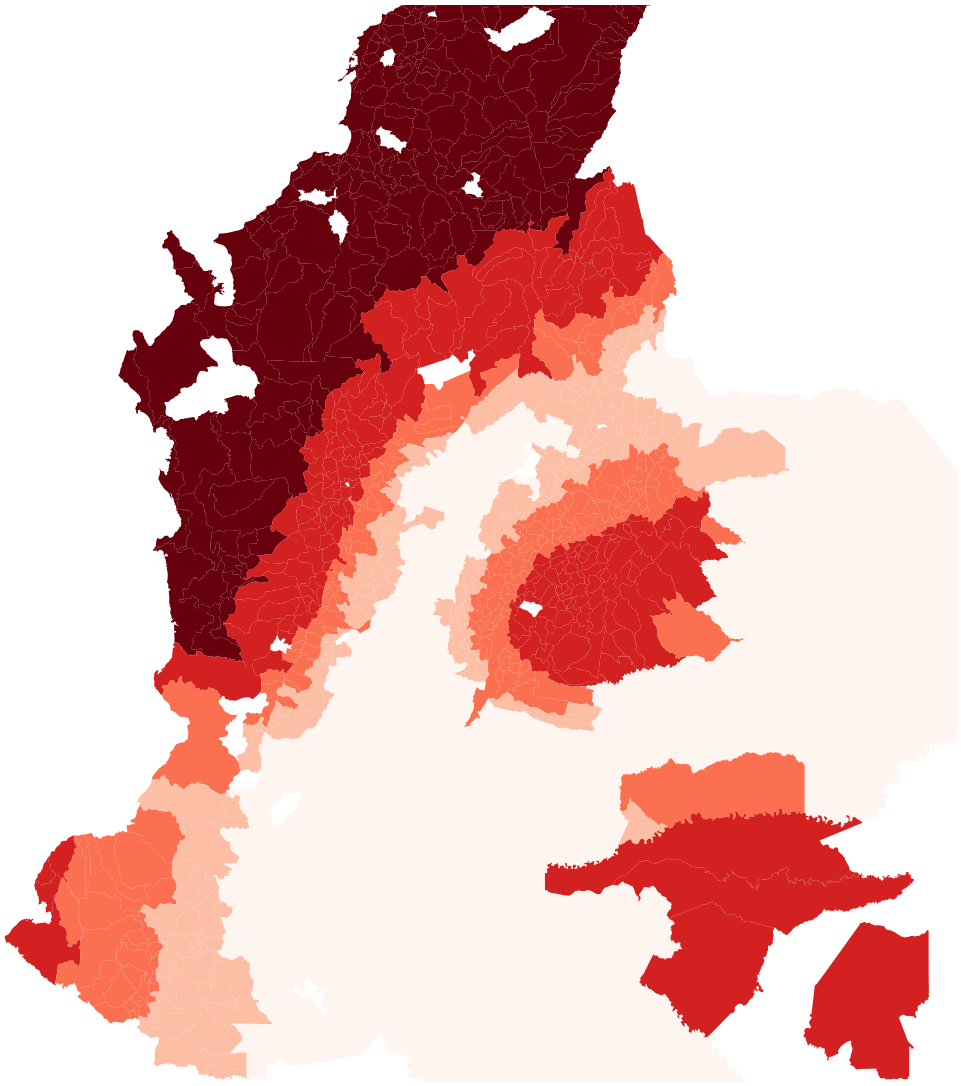

Figure C4: manuscript Figure 7

View of GWR model selection with different variables

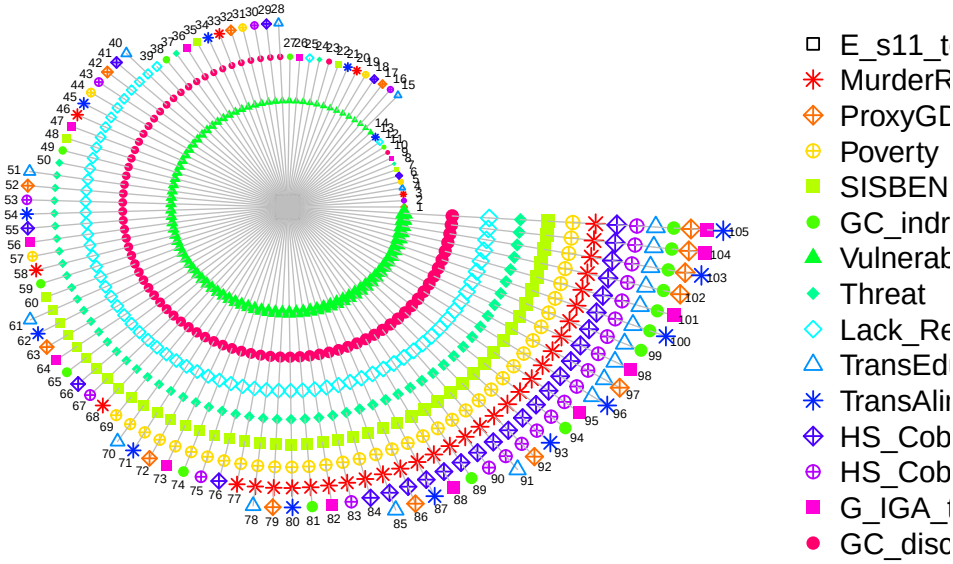

Figure C5: manuscript Figure 8

Table 2

Table 2 was reproducible. During the review, code was provided to output the data for the table, which I have saved into a CSV and rendered here (using the slightly different column headings from the CSV).

```
tab2raw <- read.csv("results/table2.csv")
keep <- c("X", "mean", "std.dev", "median", "min", "max")
tab2 <- tab2raw[,keep]
```

X	mean	std.dev	median	min	max
E_s11_to_1	48.28	3.44	48.48	36.66	59.49
SISBEN1PC	0.13	0.05	0.13	0.02	0.38
MurderRate	23.79	30.35	15.53	0.00	281.37
ProxyGDP	1.37	0.70	1.20	0.43	7.74
Poverty	89.61	21.81	90.64	14.00	163.55
GC_indrura	0.55	0.24	0.59	0.00	0.98
Vulnerabil	4.61	1.93	4.45	0.00	10.00
Threat	3.39	2.30	2.84	0.08	10.00
Lack_Respo	6.13	2.25	6.92	0.00	9.31
TransEducP	59,734.11	84,193.92	38,960.24	11,129.02	827,067.43
TransAlimE	5,679.00	3,134.36	4,984.43	499.14	24,090.09
HS_Cober_7	39.73	28.21	34.46	0.00	100.00
HS_Cober_1	57.60	29.08	57.17	0.00	100.00
G_IGA_tota	64.55	9.44	65.89	29.91	90.95
GC_discapi	78.98	56.10	66.69	0.00	376.12

Table C1: Reproduction of Table 2.

Table 3

Table 3 was reproducible. The .csv file underlying the outputs was saved into the `Replication files/Tables/` directory. This can be read into R and rendered.

```
tab3 <- read.csv("results/table3.csv")
stars <- rep("", nrow(tab3))
p <- tab3$p.value

## order important here -- do least significant first.
if (any(sig <- (p < 0.1) ))   stars[sig] <- "*"
if (any(sig <- (p < 0.05) ))   stars[sig] <- "**"
if (any(sig <- (p < 0.01) ))   stars[sig] <- "***"
tab3$significance <- stars

## rearrange the rows to match the paper
reorder <- c(2, 14, 15, 12, 5, 7, 6, 8, 4, 13, 9, 10, 11, 3, 1)

tab3_neat <- tab3[reorder, c(2, 3, 4, 7)]
```

The R-squared value at the bottom of Table 3 in the manuscript is confirmed in the .Rout file (line 541). The number of observations can be derived from 1055 d.f. with 14 variables.

There are however two minor problems with the table:

1. *Threat* has one star in the manuscript, yet it should have two stars according to the legend.

term	estimate	std.error	significance
Vulnerabil	-6.474E-01	8.082E-02	***
G_IGA_tota	6.908E-02	9.178E-03	***
TransAlimE	-2.082E-04	4.513E-05	***
GC_indrura	1.367E+00	4.869E-01	***
Threat	1.207E-01	5.724E-02	**
Poverty	4.197E-02	7.559E-03	***
SISBEN1PC	-1.773E+01	3.871E+00	***
MurderRate	-1.156E-02	2.808E-03	***
Lack_Respo	-5.360E-02	7.173E-02	
ProxyGDP	3.229E-01	1.293E-01	**
HS_Cober_7	5.764E-03	3.870E-03	
HS_Cober_1	-5.864E-03	3.409E-03	*
TransEducP	5.273E-07	1.091E-06	
GC_discapi	2.757E-04	1.494E-03	
(Intercept)	4.564E+01	9.593E-01	***

Table C2: Reproduction of Table 3.

2. *TransEducP* in the manuscript is missing the exponent; the values in Table C1 match the output from the .Rout (5.273e-07, 1.091e-06) line 532.

Acknowledgements

I thank the authors for responding to questions during the codecheck, and for providing the helpful video to recreate key steps in QGIS.

Citing this document

Stephen J. Eglen (2022). CODECHECK Certificate 2022-001. Zenodo. https://doi.org/10.5281/zenodo.6040066

About CODECHECK

This certificate confirms that the codechecker could independently reproduce the results of a computational analysis given the data and code from a third party. A CODECHECK does not check whether the original computation analysis is correct. However, as all materials required for the reproduction are freely available by following the links in this document, the reader can then study for themselves the code and data.

About this document

This document was created using R Markdown using the `codecheck` R package. `make codecheck.pdf` will regenerate the report file.

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Manjaro Linux
##
## Matrix products: default
```

```
## BLAS:   /usr/lib/libopenblasp-r0.3.19.so
## LAPACK: /usr/lib/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## other attached packages:
##  [1] readr_2.1.1          tibble_3.1.6
##  [3] xtable_1.8-4         yaml_2.2.2
##  [5] rprojroot_2.0.2      knitr_1.37
##  [7] codecheck_0.1.0.9000 parsedate_1.2.1
##  [9] R.cache_0.15.0       gh_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1  xfun_0.29         purrr_0.3.4
##  [4] vctrs_0.3.8       generics_0.1.1    htmltools_0.5.2
##  [7] utf8_1.2.2        rlang_1.0.1       R.oo_1.24.0
## [10] pillar_1.7.0      httpcode_0.3.0    glue_1.6.1
## [13] DBI_1.1.2         R.utils_2.11.0    lifecycle_1.0.1
## [16] stringr_1.4.0     R.methodsS3_1.8.1 memoise_2.0.1
## [19] evaluate_0.14     tzdb_0.2.0        fastmap_1.1.0
## [22] curl_4.3.2        fansi_1.0.2       highr_0.9
## [25] osfr_0.2.8        cachem_1.0.6      rorcid_0.7.0
## [28] jsonlite_1.7.3    fs_1.5.2          hms_1.1.1
## [31] digest_0.6.29     stringi_1.7.6     dplyr_1.0.7
## [34] cli_3.1.1         tools_4.1.2       magrittr_2.0.2
## [37] crul_1.2.0        crayon_1.4.2      whisker_0.4
## [40] pkgconfig_2.0.3   ellipsis_0.3.2    fauxpas_0.5.0
## [43] assertthat_0.2.1  rmarkdown_2.11    httr_1.4.2
## [46] R6_2.5.1          compiler_4.1.2
```