Which is Bigger: Switzerland or Chad? Modeling Size Variation in Country Embeddings

Sara Bartl, Jack Grieve, Jason Grafmiller, Akira Murakami
University of Birmingham, UK

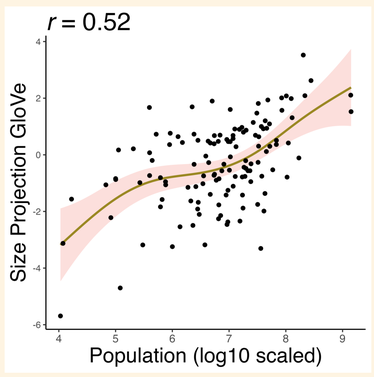
RQ: What external factors affect a country's relative size in the GloVe embedding space?

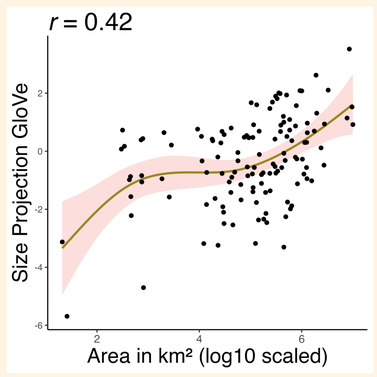
Methods

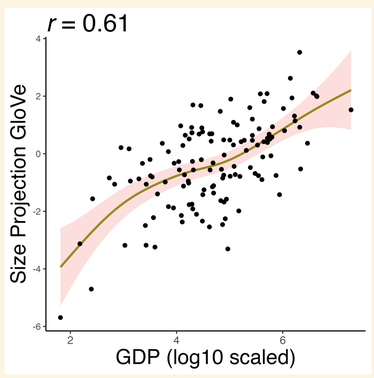
- obtain countries' relative size in the GloVe embedding space through semantic projection
- build a linear regression model with 3 features: population, surface area and GDP
- analyse residuals for patterns

Correlation Analysis

	pop_log	area_log	gdp_log	projection
pop_log	1.000	0.840	0.815	0.519
area_log	0.840	1.000	0.668	0.423
gdp_log	0.815	0.668	1.000	0.608
projection	0.519	0.423	0.608	1.000

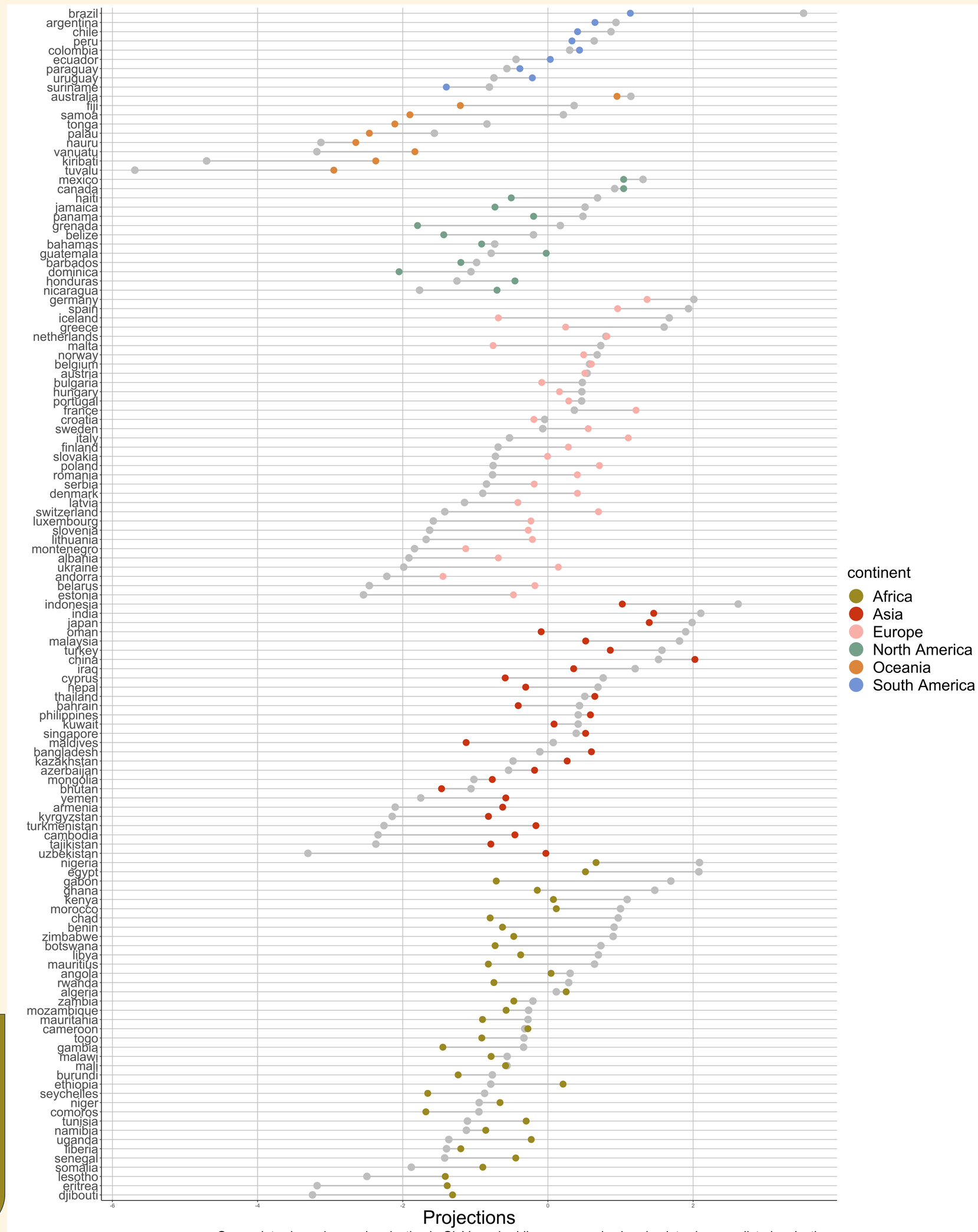






GDP is an important factor in accounting for the variation in the GloVe embeddings. Population and Area are also meaningful.

Variation in Model 1 Residuals by Continent



Grey points show observed projection in GloVe embedding space and colored points show predicted projections

Linear Regression Model 1

Parameter	Coefficient	SE	95% CI	t(125)	p
(Intercept)	-0.31	0.11	[-0.52, -0.10]	-2.94	0.004
gdp log c	0.81	0.18	[0.45, 1.16]	4.49	< .001
population log c	0.13	0.27	[-0.40, 0.66]	0.49	0.625
area log c	-0.02	0.17	[-0.36, 0.32]	-0.11	0.912

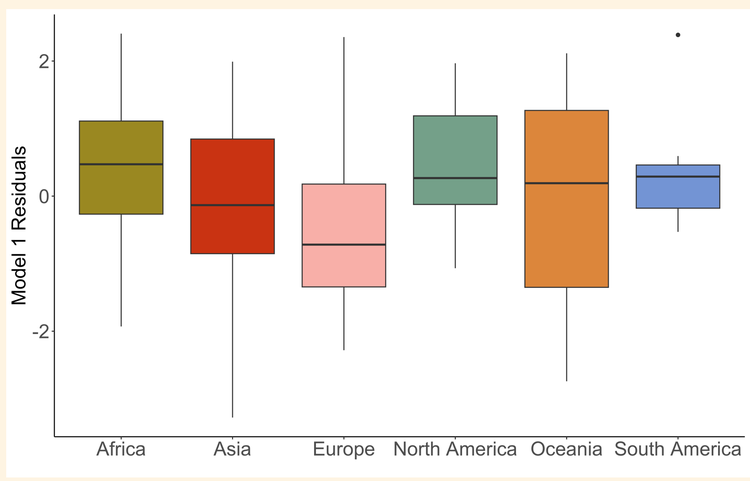
Adjusted R-squared: 0.3568

Linear Regression Model 2

Parameter	Coefficient	SE	95% CI	t(120)	p
(Intercept)	-0.27	0.12	[-0.51, -0.03]	-2.27	0.025
africa	0.81	0.24	[0.33, 1.29]	3.34	0.001
asia	-0.18	0.23	[-0.63, 0.27]	-0.79	0.431
europe	-1.08	0.24	[-1.56, -0.59]	-4.39	< .001
north america	0.28	0.28	[-0.27, 0.84]	1.01	0.316
oceania	-0.06	0.38	[-0.81, 0.68]	-0.17	0.868
gdp log c	1.63	0.24	[1.15, 2.11]	6.69	< .001
population log c	-0.53	0.30	[-1.13, 0.07]	-1.76	0.081
area log c	-0.14	0.17	[-0.47, 0.19]	-0.82	0.412

Adjusted R-squared: 0.4484

Residual Variation



Contact: s.bartl@bham.ac.uk

Scan for data and code

We find a potential continent bias in countries' relative size in the embedding space.

African countries are bigger in the GloVe embeddings than predicted by the regression model. European countries on the other hand appear smaller in the embeddings than predicted.