Clustering

1 Open issues

- What about the source/greeness of energy production? These are also in the energy balance data set
  - **Suggestion**: For total domestic energy supply export of energy only use energy that has been produced by green sources?
- Consideration of employment dimension
- Consideration of green products in export basket
- Discuss results against the descriptives for the clusters below, so far the discrimination is not very convincing

2 Data setup

We focus on the time between 2014 and 2018 and consider the following variables:

Dimension	Rationale	Variable	Source
Externalization of ecological stressors	Countries that externalize more have a less sustainable more of provisioning.	Net GWP imports per capita	EXIOBASE, own calculation.
Total domestic GHG emissions	Countries the emit a lot GHG need to change more drastically.	GWP emissions per capita	EXIOBASE, own calculation.
Economic strength	Countries with more economic strength can more easily support the transition and counter challenges.	Domestic value added	EXIOBASE, own calculation.
Domestic energy supply	Countries with high domestic supply **IF GREEN?** are more independent and more flexible.	Total primary energy production per capita	Eurostat

Dimension	Rationale	Variable	Source
Domestic energy demand	Countries with high domestic demand are less flexible and more dependent.	Total energy consumption per capita	Eurostat
Export of energy	Countries that export more energy **IF GREEN?** are more likely to benefit.	Total energy exports per capita	Eurostat
Innovation in green technologies	Countries with technological capabilities in green areas are likely to benefit more.	Green patents per million people (EPO classification)	PATSTAT
Dependence on brown employment	Countries that depend a lot on brown jobs face bigger challenge of transition.	**TBD**	
Production of green products	Countries that produce many green products are likely to benefit more.	**TBD**	

3 Conduct the clustering

3.1 Choice of the clustering algorithm

Data preparation:

There are four different agglomerative cluster algorithm. Usually one chooses the one that yields the highest clustering coefficient:

Algorithm	Coefficient
average	0.7835744
single	0.7690382
complete	0.8072610
ward	0.8249842

Thus, in our case we should use the WARD algorithm.

We might also assess the quality by comparing the cophenetic distance of the clustered data and the original euclidean distances:

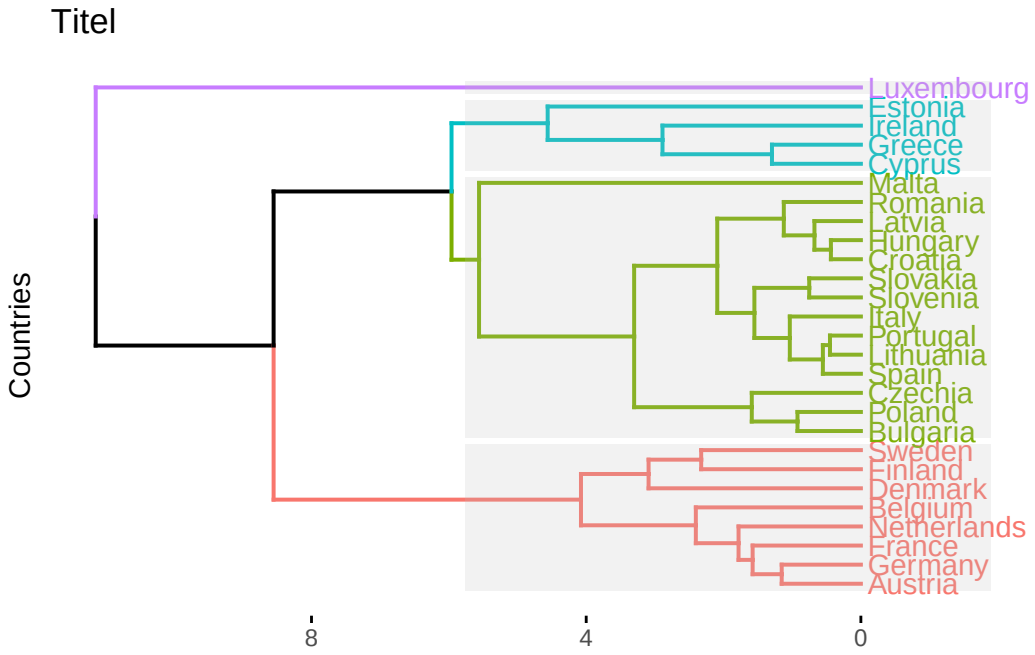
```
[1] 0.7238995
```

This is a satisfactory number.

What remains open is the right number of clusters, but this is also a question of interpretation.

3.2 Result of the clustering

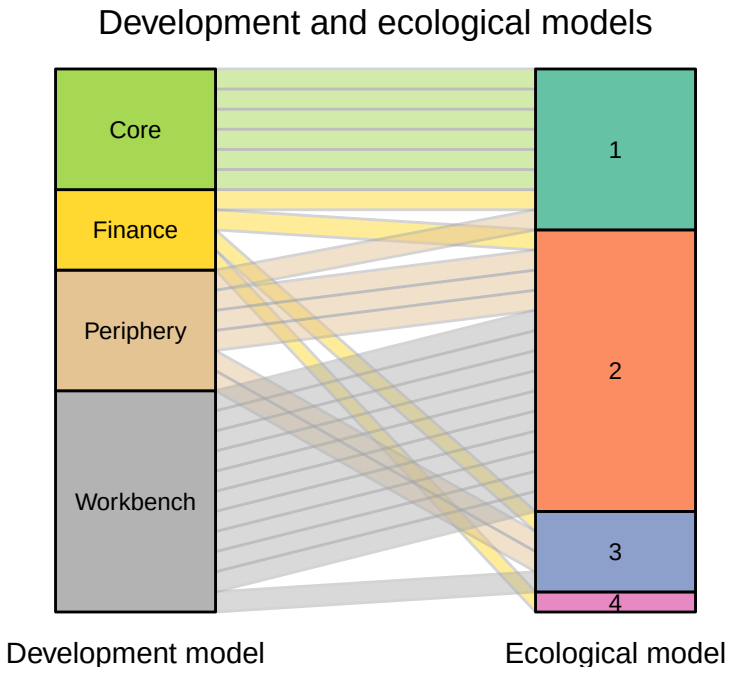
Here is the overall result of the clustering assuming 4 clusters.



Ecological model 1	Ecological model 2	Ecological model 3	Ecological model 4
Austria	Bulgaria	Cyprus	Luxembourg
Belgium	Czechia	Estonia	
Germany	Spain	Greece	
Denmark	Croatia	Ireland	
Finland	Hungary		
France	Italy		
Netherlands	Lithuania		
Sweden	Latvia		
	Malta		
	Poland		
	Portugal		
	Romania		
	Slovenia		
	Slovakia		

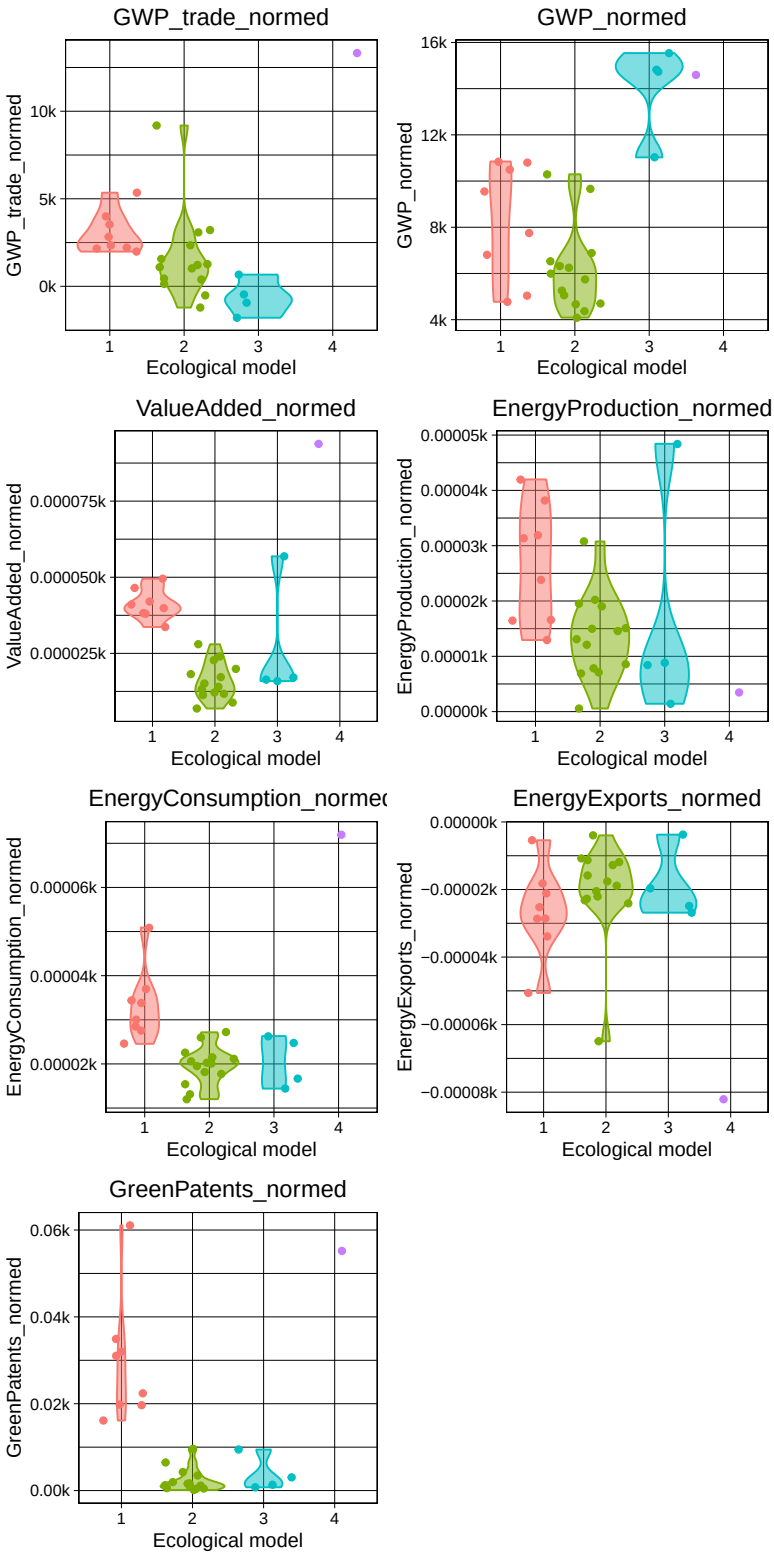
Ecological model 1	Ecological model 2	Ecological model 3	Ecological model 4

Here I visualize the relationship to the development models classification:



3.3 Descriptive statistics for the clusters

This is not weighted by population. Should we do this?


GWP_trade_normed
GWP_trade_normed
10k
5k
0k
1
2
3
4
Ecological model
GWP_normed
GWP_normed
16k
12k
8k
4k
1
2
3
4
Ecological model
ValueAdded_normed
ValueAdded_normed
0.000075k
0.000050k
0.000025k
1
2
3
4
Ecological model
EnergyProduction_normed
EnergyProduction_normed
0.00005k
0.00004k
0.00003k
0.00002k
0.00001k
0.00000k
1
2
3
4
Ecological model
EnergyConsumption_normed
EnergyConsumption_normed
0.00006k
0.00004k
0.00002k
1
2
3
4
Ecological model
EnergyExports_normed
EnergyExports_normed
0.00000k
−0.00002k
−0.00004k
−0.00006k
−0.00008k
1
2
3
4
Ecological model
GreenPatents_normed
GreenPatents_normed
0.06k
0.04k
0.02k
0.00k
1
2
3
4
Ecological model