The variable selection in the regional models was initially following the same conceptual guidelines as the global model, but due to the higher number of observation and different aggregation of the data, we have made an exploratory data analysis to verify that there are no hidden correlations or dependencies in the data we have not accounted for.

Firstly, we have tested for spatial autocorrelation (Moran’s test), where we found a weak spatial autocorreltion on a relatively low signficance level.

Second, we used a ‘brute force’ approach to run all possible linear regressions and multiple regressions on all available data, to see if there is any spurious or real connection among the variables. The code exectuing this exploratory analysis is in the supporting repository.

Third, we applied a more systematic variable selection method, by creating a random forest of regression models. We did not intend to choose a model candidate from the random forest regressions to test or hypothesises, mainly, because machine learning models may have high predictive power, but they are often ‘black boxes’ that make conceptual explanations difficult, and their results are not directly comparable with the global model. Our ‘brute force’ approach was comprehensive in the way that it has measured all possible regression models, and the random forest comprehensively compared the predictive power of variables groups. Given that both methods are comprehensive and did not reveal any further reserach directions, we found our initial theoretical framework validated, and based on this variable selection we created comparable regional models to the global model.