Recostructing climate change awareness survey data using socio-economics and geographical information

Gianpaolo Zammarchi

Abstract Spatially clustered regression is a recently developed methodology that combines clustering and regression in order to provide region-specific predictions. In this work, we applied spatially clustered regression to predict climate change awareness in countries not included in a survey conducted across approximately 110 countries or world areas. We are using socio-economics and geographical information to train models capable of estimating awareness after assessing the underlying relationships at regional level. Estimating a global level of awareness about climate change is crucial to assess the knowledge of this problem across the world before policies, collaborations and plans can be effectively implemented.

Key words: spatially clustered regression, climate change, awareness imputation

1 Introduction

Linear regression is a simple yet powerful tool for analyzing the relationship between a dependent variable and one or more independent variables. In this work we will focus on cluster regression and specifically on spatially clustered regression (SCR).

Generally speaking, cluster regression is a statistical modeling technique that takes into consideration the existence of groups of homogeneous observations in the data. The relationship between the dependent and independent variables may differ according to the subgroups to which the data points belong to. Finding these groups and estimating distinct regression models for each group is the aim of cluster regression. SCR is a method that combines regression models with clustering, under the assumption all the geographical locations that are close in space can share the

Gianpaolo Zammarchi
Department of Economics and Business Sciences, Viale Fra Ignazio, 17, 09131, Cagliari, Italy, e-mail: gianpaolo.zammarchi@unica.it

same cluster. Therefore, the regression coefficients obtained for a cluster are valid for all countries in the same cluster [1].

In our study, we decided to use a large-scale survey promoted by Yale researchers in collaboration with Meta that originally provided data for 110 countries in the world [2]. Unfortunately, some countries were not involved in the survey and therefore the levels of awareness were not available. However, it is possible to use SCR to estimate the potential awareness values for countries not involved in the survey. Assessing and possibly monitoring people's levels of awareness about climate change is of fundamental importance to continue to pressure governments and supranational organizations to adopt measures to reduce greenhouse gas emissions and encourage virtuous behavior also by individuals. Since awareness levels for some countries might not be available for several years, the aim of this work is to provide reliable estimates of what these values could be in order to understand the current situation within these countries.

2 Background

SCR divides data into clusters based on shared characteristics in order to create regression models that account for spatial heterogeneity. Several approaches have been proposed in the literature to address the challenge of capturing spatial heterogeneity by estimating local models on subsets of data centered around a specific location. For instance, Lee and colleagues [3] proposed a method based on geographically weighted regression and applied it to a cancer mortality data set.

Li & Sang [4] proposed a computationally efficient method for detecting spatially clustered patterns in regression coefficients using a combination of spatial neighborhood information and regularization techniques.

Lastly, Fu and colleagues [5] used a smoothing method to obtain the asymptotic covariance matrices of the parameter estimators and prove the asymptotically unbiasedness of the induced estimating functions. Their approach consisted in the creation of covariance matrices for both between cluster estimators and within cluster estimators, along with a combined estimator of the within-cluster correlations. When applied to a data set of randomized clinical trials, the method showed good computational efficiency and stable results.

3 Methodology

In this section we illustrate how we created the data set for the analysis and provide an explanation of the methods used to carry out these analysis.

3.1 Data set

The awareness levels were obtained by collecting the answers given to a survey administered in 2022 by researchers at Yale University in collaboration with Meta [2]. The outcome variable was obtained as the number of respondents who answered positively to the question on awareness of climate change in the survey. The socio-economics and geographical information used as predictors in the regression models were collected from the following sources: Our World in Data, Penn World Table, Global Carbon Atlas, and Climate Change Knowledge Portal.

3.2 Spatially clustered regression

SCR is a modeling approach that finds clusters in geographically-related data, where the relationship between predictors and the response variable is assumed to be the same within each cluster. The main goal is to fit different regression models for each of the K clusters. Notationally, consider a response variable Y_i and p predictors $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ observed at location i. In SCR's formulation the relationship between Y_i and $\mathbf{X}_i$ is the same as that of a linear regression model:

$$Y_i = \mathbf{X}_i^\top \beta_{C(i)} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

where $C(i)$ indicates the cluster for location i, β is the vector of regression coefficients specific of cluster $C(i)$, and ε_i is the error term. The clustering is obtained by minimizing an objective function that combines within-cluster error and spatial heterogeneity as in Equation 1:

$$\mathscr{L} = \sum_{k=1}^{K} \sum_{i \in C(i)} (Y_i - \mathbf{X}_i^\top \beta_{C(i)})^2 + \phi \sum_{i<j} w_{ij} \mathbb{I}[C(i) = C(j)], \tag{1}$$

where ϕ is a parameter controlling for the strength of spatial correlation.

4 Results

The analysis is still in a preliminary phase. Figure 1 shows a clustering of the countries based on the climate change awareness level. A three cluster solution shows a strong effect of geographical proximity, with a purple cluster mainly including European countries, a green cluster including South American and Asian countries and a yellow cluster mostly including South American and African countries. SCR analysis is ongoing and will allow to compare clusters obtained using both geographically-related and socioeconomic predictors, as well as to obtain country-specific predictions.

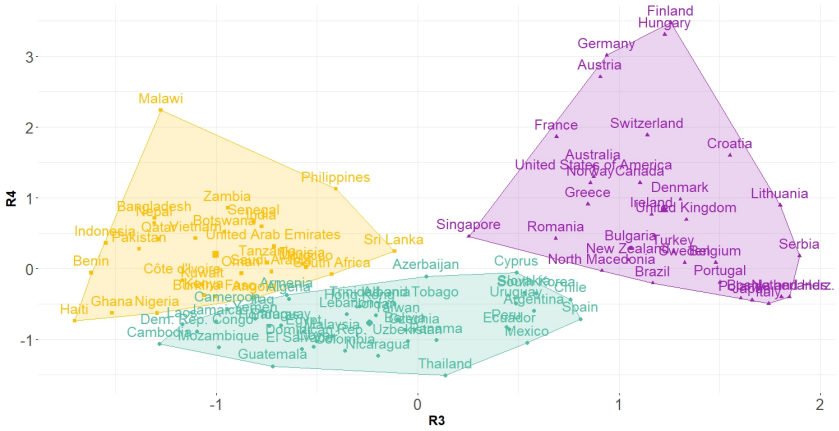

Fig. 1 Countries clustered based on their awareness levels

References

1. Sugasawa, S., & Murakami, D. (2021). Spatially clustered regression. Spatial Statistics, 44, 100525.
2. Leiserowitz, A., Carman, J., Buttermore, N., Wang, X., Rosenthal, S., Marlon, J., & Mulcahy, K. (2021). International public opinion on climate change. New Haven, CT: Yale Program on Climate Change Communication and Facebook Data for Good.
3. Lee, J., Gangnon, R. E., & Zhu, J. (2017). Cluster detection of spatial regression coefficients. Statistics in medicine, 36(7), 1118-1133.
4. Li, F., & Sang, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. Journal of the American Statistical Association.
5. Fu, L., Wang, Y. G., & Bai, Z. (2010). Rank regression for analysis of clustered data: a natural induced smoothing approach. Computational statistics & data analysis, 54(4), 1036-1050.