ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

Comparison of Different Methodological Approaches to Simulate Geo-Referenced Populations to Be Used in a Cluster Analysis of Childhood Leukaemia Cases in Germany

Di Staso Rossana[1,2], *Gianicolo Emilio*[2]

(1) *Department of Medical and Surgical Sciences, University of Bologna (Italy)*
(2) *Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), University of Mainz (Germany)*

CORRESPONDING AUTHOR: *Di Staso Rossana, rossana.distaso2@unibo.it*

INTRODUCTION

A key methodological challenge in epidemiological studies using a cluster analysis approach is the choice of an appropriate set of controls. This challenge becomes particularly complex when cases are geo-referenced and the outcome is rare. In fact, in such situations, controls need to be sampled from a comprehensive primary base, where population is defined both geographically and temporally. Furthermore, if cases are geo-referenced, the controls need to be geo-referenced too. However, selecting and geo-referencing such controls can be highly resource-intensive, both in terms of time and cost.

OBJECTIVE

Thus, the main objective of our study is to use publicly available data and established geo-statistical techniques to simulate a geo-referenced population (GRP). This simulated geo-referenced population will be then used as the primary basis for the extraction of controls in a cluster analysis that will focus on childhood leukaemia incident cases in Germany.

METHODS

For the period 2000-2020, we used population counts of persons aged 0 to 14 years from the WorldPop's (WP) project at the University of Southampton and available in 100×100m grid cells [1]. The WP project employs a top-down modelling approach and uses different types of variables (rural settlements, industrial areas, schools, etc.) to estimate age-specific (0, 1–4, 5–9, 10-14 years) and sex-specific counts of persons in a grid [2,3,4]. The observed population figures (RP) at the municipality level were provided by the German Childhood Cancer Registry and were used as constraint values. To simulate a georeferenced population, the WP was used as a probability distribution function for sampling, with replacement, a number of cells equal to the RP. Afterwards, a uniform distribution was applied to randomly sample inside each picked cell a number of points (coordinates) equal to the times the cell was extracted.

Here are shown results for three years: 2004, 2011 and 2019 and three simulated GRPs. Whereby the three simulations refer to the geographical level used to constrain the simulated population to the real population, i.e. the overall Childhood German population (S1), the childhood population at the state level (Bundesland) (S2), and the population at province level (Landkreis). For evaluation purposes, the percentage differences between the SP and the RP for all German municipalities were computed and summarized as the median and interquartile range (IQR). In addition, the root mean squared error (RMSE) was calculated.

RESULTS

In Germany, the number of children between 0 and 14 years old was 12,045,019 in 2004, 10,832,081 in 2011, and

DOI: 10.54103/2282-0930/29370

11,396,196 in 2019. The WP estimations for the same years were 12,178,611, 10,886,770, and 10,457,921, respectively. When using the overall childhood population of Germany as the constrain for the simulation (S1), in 2004 we observe an overestimation of the population in the eastern Germany and an underestimation elsewhere (Figure 1; a) (median percentage difference =-7.1; IQR: -13.3 – 7.8); RMSE of 17.8. Percentage differences decreases in the third simulation (S3: median= -3.5; IQR: -12.2 – 5.1; a RMSE=6.5). Simulations for 2011 show, in general, better results with S3 as best performance (median=-1.2; IQR: -10.1 – 7.3; RMSE 4.4). Generally, results observed in 2019 are similar to those observed in 2011.

CONCLUSIONS

Despite being computationally the most time-consuming, S3 shows the best performances in terms of narrower interquartile ranges and a more centered distribution. Thus, the simulated georeferenced population obtained using the RP at the province level as the constrain can be considered the optimal one. Further investigations are needed to shed light on the geographical differences observed in 2004.

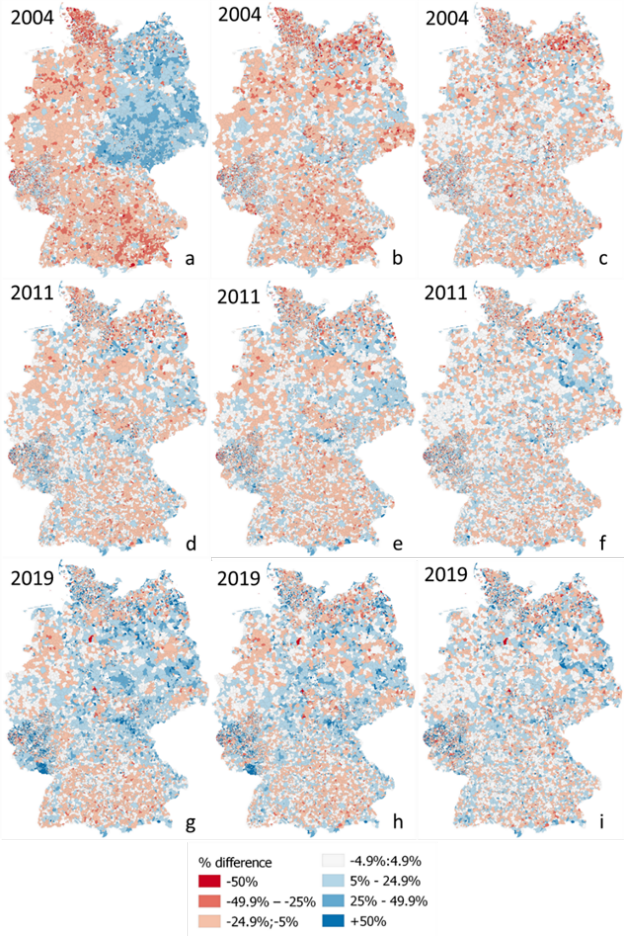
The inconsistencies between the WP and the RP must be considered as a limitation when interpreting our results. However, this is an innovative method which allows the future use of the overall georeferenced population or a selection of it for cluster analyses.

The application of this method to each year of interest and the cluster analysis itself are pending.

REFERENCES

1. Tatem, A. J. WorldPop, open data for spatial demography. Sci. Data 4, 2017;170004
2. Stevens FR, Gaughan AE, Linard C. et al., Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE, 2015;10(2):e0107042.
3. Sorichetta, A., Hornby, G., Stevens, F. et al. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Sci Data 2, 2015, 150045
4. Gaughan, A., Stevens, F., Huang, Z. et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010. Sci Data 3, 2016, 160005

Figure 1. Percentage difference between the Population given by the registry and the GRP S1, GRP S2 and GRP S3 in 2004 (a, b and c), in 2011 (d, e and f) and in 2019 (g, h and i).



DOI: 10.54103/2282-0930/29370