MethodsX 8 (2021) 101257

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Small-scale spatial distribution of COVID-19-related excess mortality

Dino Gibertoni[a,*], Francesco Sanmarchi[b], Kadjo Yves Cedric Adja[b], Davide Golinelli[a], Chiara Reno[b], Luca Regazzi[c], Jacopo Lenzi[a]

[a] *Unit of Biostatistics and Hygiene, Department of Biomedical and Neuromotor Sciences, University of Bologna, Italy*
[b] *Specialty School of Hygiene and Preventive Medicine, University of Bologna, Italy*
[c] *University of Bologna, Italy*

ABSTRACT

Mortality due to massive events like the COVID-19 pandemic is underestimated because of several reasons, among which the impossibility to track all positive cases and the inadequacy of coding systems are presumably the most relevant. Therefore, the most affordable method to estimate COVID-19-related mortality is excess mortality (EM). Very often, though, EM is calculated on large spatial units that may entail different EM patterns and without stratifying deaths by age or sex, while, especially in the case of epidemics, it is important to identify the areas that suffered a higher death toll or that were spared. We developed the Stata COVID19_EM.ado procedure that estimates EM within municipalities in six subgroups defined by sex and age class using official data provided by ISTAT (Italian National Statistics Bureau) on deaths occurred from 2015 to 2020. Using simple linear regression models, we estimated the mortality trend in each age-and-sex subgroup and obtained the expected deaths of 2020 by extrapolating the linear trend. The results are then displayed using choropleth maps. Subsequently, local autocorrelation maps, which allow to appreciate the presence of local clusters of high or low EM, may be obtained using an R procedure that we developed.

- We focused on estimating excess mortality in small-scale spatial units (municipalities) and in population strata defined by age and sex.
- This method gives a deeper insight on excess mortality than summary figures at regional or national level, enabling to identify the local areas that suffered the most and the high-risk population subgroups within them.
- This type of analysis could be replicated on different time frames, which might correspond to successive epidemic waves, as well as to periods in which containment measures were enforced and for different age classes; moreover, it could be applied in every instance of mortality crisis within a region or a country.

DOI of original article: 10.1016/j.healthplace.2021.102508

[*] Corresponding author.
E-mail address: dino.gibertoni2@unibo.it (D. Gibertoni).

https://doi.org/10.1016/j.mex.2021.101257

ARTICLE INFO
Method name: Excess mortality estimation
Keywords: Excess mortality, Municipalities, COVID-19 diffusion, Spatial analysis
Article history: Received 20 November 2020; Accepted 29 January 2021; Available online 3 February 2021

Specifications table

Subject Area:	Medicine and Dentistry
More specific subject area:	Epidemiology, Public Health
Method name:	Excess mortality estimation
Name and reference of original method:	Not applicable
Resource availability:	Official data on daily mortality of a country or region(s) at the municipality level, covering several years; geographical data of the same country or region(s).

Method details

Several methods have been used to estimate excess mortality (EM), under the common basic idea to capture the normal magnitude of deaths in a non-perturbed time period (the observation period) and project the same magnitude on the period in which the mortality crisis happened. What differentiates these methods is the analytical tool that is used to estimate the underlying magnitude of deaths. The simplest ideas employ the yearly average of deaths over the observation period, or the moving average on a convenient number of years, while more advanced analytical tools include difference-in-difference analysis, Poisson or negative binomial regression, and time series analyses. As is always the case, advanced tools tend to provide more accurate estimates in that they can effectively model potential time trends or adjust for potential confounders; however, they require higher-quality data and sometimes a higher educated public. Conversely, simpler tools can be applied when only basic data are available and are usually understood by wide potential audiences. The choice between these methods is then dictated by striking the balance between the quality of available data, the desired level of insight and the main audience target to which the study is addressed.

However, what we often see is that even when abundant and affordable data are available, EM estimates are produced only at a basic and concise level, that is providing only summary national or regional figures. Regardless of the analytical tool used for the estimation, this is a major limitation that precludes any chance of identifying local areas that were the epicenters of the pandemic or, on the other hand, that were spared by the pandemic. An equally serious limitation is the absence of population stratification, which is essential to understand which subgroups are at higher and lower risk, which could coexist even within the most hit areas.

To overcome these limitations, we developed a procedure that provides EM ratios calculated at the small-scale level (municipalities) in age-and-sex population subgroups. This procedure can be applied in all instances when official data on mortality are provided with the required granularity, as is the case in Italy where the data source is ISTAT, and using the Stata and R scripts that we include in this paper as supplementary materials. Specifically, Stata was used for EM estimation and to obtain choropleth maps of EM, while R was used for spatial analysis.

Datasets and data preparation

ISTAT releases periodical updates of mortality data, covering an increasing period from January 1st to a fixed endpoint for each year from 2015 to 2020 [4]. These data include the daily number of deaths occurred in every Italian municipality that provided complete data, grouped by sex and age classes. To prepare the data for the estimation of EM, a selection was made in order to obtain datasets specific for males and females of each region, covering a predetermined time period. The time frame of interest for the analysis must also be selected before running the program. Grouping by age was also necessary; specifically, we created three classes appropriate for the analysis of COVID-19

Table 1
Description of the dataset to be used for EM estimation.

Variable name	Storage type	Display format	Value label	Variable label
mun_code	long	%8.0g		Code of municipality
reg_code	byte	%8.0g		Code of region
Region	str29	%29s		Region name
Province	str29	%29s		Province name
Municipality	str34	%34s		Municipality name
age_class	byte	%8.0g	age_class	Age class
M_15	Int	%8.0g		Male deaths in 2015
M_16	Int	%8.0g		Male deaths in 2016
M_17	int	%8.0g		Male deaths in 2017
M_18	int	%8.0g		Male deaths in 2018
M_19	int	%8.0g		Male deaths in 2019
M_20	int	%8.0g		Male deaths in 2020
F_15	int	%8.0g		Female deaths in 2015
F_16	int	%8.0g		Female deaths in 2016
F_17	int	%8.0g		Female deaths in 2017
F_18	int	%8.0g		Female deaths in 2018
F_19	int	%8.0g		Female deaths in 2019
F_20	int	%8.0g		Female deaths in 2020

mortality: 0 to 64 years, 65 to 74 years, and 75 years or more (cutoffs may be modified by users according to their convenience). The dataset structure is reported in Table 1: municipality and region codes are included, as well as their text description and, when present, the name of the province (in Italy, province is an intermediate government level amidst municipality and region). Among these variables, mun_code is the only one required by EM estimation, while reg_code is required for naming the choropleth maps and the dataset that will include the estimations; Region will be used in the title of the choropleth map, while Province and Municipality are kept only to assist in data checking. The age_class variable was obtained by recoding the original age classification provided by ISTAT, as previously described. M_15 to M_20 are variables reporting the number of deaths for males in years from 2015 (M_15) to 2020 (M_20), and F_15 to F_20 are the variables including the corresponding number of deaths for females. These variables are obtained by summing the daily numbers of deaths over the observation period of interest.

Excess mortality estimation

The estimation of EM is performed by the stand-alone Stata procedure that we generated, called COVID19_EM.ado, which loads the sets of summarized mortality data and in sequence estimates EM and draws regional choropleth maps. This procedure must be run from the program line of Stata and requires two arguments: the ending date of the period of interest and sex, with the region codes as optional argument. Thus, for instance, to obtain the maps of EM up to April 30th, 2020 for females (F) in Lombardy (region code 3), one must type the following command:

```
COVID19_EM 300420 F, region(3)
```

To analyze more than one region, the user can specify all the required codes in the optional argument. For instance, to obtain the previous EM and maps for Lombardy, Veneto and Emilia-Romagna we typed:

```
COVID19_EM 300420 F, region(3 5 8)
```

For consecutively coded regions, a convenient Stata notation can be used, in the following example to obtain estimates and maps for all regions coded from 3 to 8:

```
COVID19_EM 300420 F, region(3(1)8)
```

When the region() option is not specified, the analysis is run on the whole of Italy, that is on the whole dataset with mortality data.

The expected number of deaths for each age-and-sex subgroup in each municipality was estimated by linear regression using the total number of deaths (D) as dependent variable and the ordinal number of the year (Year) as independent variable; Year=0 was assigned to 2015 and each subsequent year increased by 1 up to Year=4 for 2019.

$$D = \alpha + \beta * Year$$

To predict the expected mortality in 2020, we extrapolated the estimated linear trend for each municipality/age group/sex, using the equation:

$$D_{2020} = \alpha + \beta * 5$$

where Year =5 corresponds to 2020 in the years' progression.

This was accomplished in the procedure by using the parmby module [6], which computes linear regression for each instance of age/sex within each municipality and then saves the results in the EMestimates.dta dataset. Observed-to-expected mortality ratio is then easily computed by a simple division of the two variables, but two further points need clarification.

First, using extremely granular data often entails the presence of very small numbers, because smaller municipalities may have few or even no deaths in a specified period of the year. The following adjustments are made in the case of negative estimates of expected mortality and in cases in which the expected deaths are zero, or observed deaths are zero, or both, as follows:

- If expected deaths are =0 and observed deaths are >0, then EM equals observed mortality. For instance, if there are 0 expected deaths and 2 observed deaths, then EM=200%
- If observed deaths are =0, and expected deaths are =0 or >0, then EM = 0. For instance, if expected deaths are 2 and observed deaths are 0, then EM = 0%
- If expected deaths are <0, then we considered them as =0 and proceeded according to the previous points.

Second, as the period of interest may include the month of February, and particularly when EM is estimated for a leap year (which is the case when examining COVID-19-related EM for 2020), a correction to the expected number of deaths to account for the higher number of days in the leap year is recommended. This was done by counting the number of days in the leap year over which EM is estimated, and considering that in the previous five years only one of them was a leap year, so that their number of days is 0.8 less than those of the leap year. To put it simply, while February 2020 had 29 days, the average number of days in February 2015–2019 was 28.2. Thus, the correction factor that we used was given by this ratio:

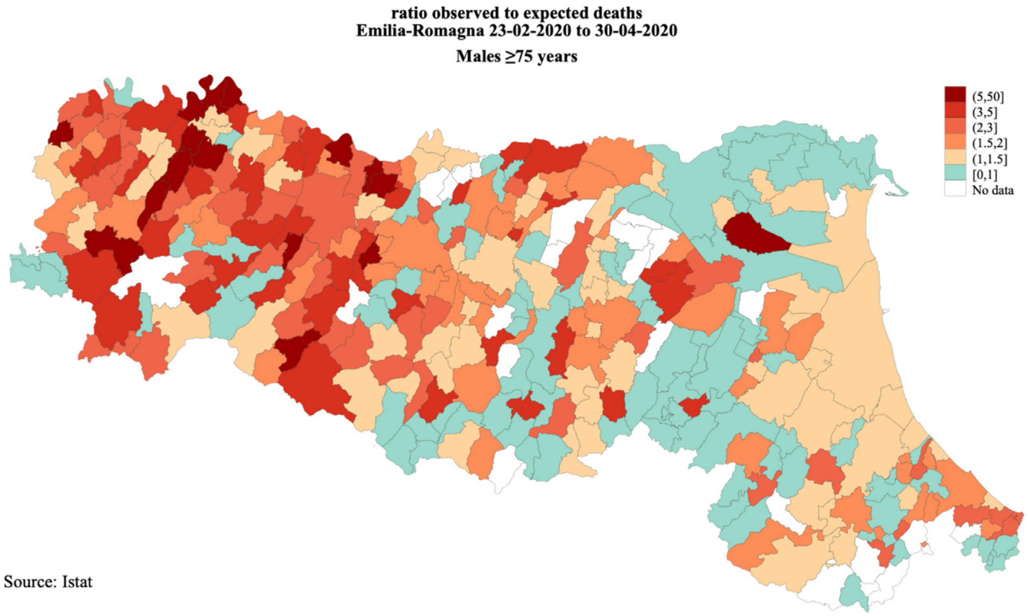
$$\frac{n.\ of\ days\ in\ 2020}{n.\ of\ days\ in\ 2020 - 0.8}$$

Obviously, this correction factor becomes more and more negligible as the width of the period of interest for EM estimation increases.

Another section of the procedure was necessary to address the cases of small municipalities for which, under one or more age classes, no deaths occurred. To correctly draw the choropleth maps, all spatial units are always required, therefore the database needs to be expanded by the number of lines corresponding to the number of missing age classes, and this is done in lines 34–37 and 132–136 of the script.

Two different outputs are provided by the COVID19_EM procedure: a Stata dataset including all EM estimates and choropleth maps for each age class in the selected region and sex. If the analysis was required for more regions, datasets and maps are produced for each region separately. To draw choropleth maps, users must get, possibly from official sources, the datasets including information on centroids and polygon coordinates for the examined spatial units. These datasets must be merged to the dataset including the variable to be drawn in maps (which in our case is estratio, i.e. the observed-to-expected mortality ratio). Then, the user-written spmap procedure [10] is used to draw maps, using a customized color pattern that assigns light green to municipalities with ratio ≤1 (those with negative EM) and five increasing hues of red (defined by the fcolor option) as the ratio shows values >1 at five different thresholds (defined by the clbreaks option). An example of the maps which may be obtained is given in Fig. 1.

The COVID19_EM.ado procedure is reported below in the Supplementary materials, so that those interested can use it and if necessary modify its settings. In particular, the time period in which EM


ratio observed to expected deaths
Emilia-Romagna 23-02-2020 to 30-04-2020
Males ≥75 years
(5,50]
(3,5]
(2,3]
(1.5,2]
(1,1.5]
[0,1]
No data

Source: Istat

Fig. 1. Excess mortality of 2020 with respect to 2015-19 in the municipalities of Emilia-Romagna, in the period February 23rd - April 30th and for males aged ≥ 75 years.

is calculated has a fixed starting date set to February 23rd, 2020, which we identified as the starting date of the COVID-19 pandemic in Italy. Thus, with the previous command we estimated EM using all deaths occurred between February 23rd and April 30th among females living in Lombardy, Veneto and Emilia-Romagna. The starting date can be modified by the user, for instance to obtain EM on a monthly basis, and feeding the command with the appropriate dataset.

Spatial autocorrelation

Spatial autocorrelation refers to the dependencies that exist among observations that are proximal within geographic space. These dependencies produce clustering of similar (positive spatial autocorrelation) or dissimilar (negative spatial autocorrelation) values, and hence produce some “map pattern”. Classic spatial autocorrelation statistics include Moran’s I and Getis–Ord general G [3,5], which estimate the overall degree of spatial autocorrelation for a dataset. This is a fundamental piece of information, especially as a diagnostic tool for spatial model misspecification. However, the degree of autocorrelation may vary significantly across geographic space. Local spatial autocorrelation statistics provide estimates to the level of the spatial unit of analysis, allowing for some visual exploration of map patterns and for some discrimination between local areas according to their degree of dependency. Local versions of I (local Moran’s I_i) and G (Getis–Ord G_i^*) are both available [1,7]

Local Moran’s I_i and Getis–Ord G_i^* capture different aspects of local autocorrelation. Local Moran’s I_i is given by:

$$I_i = \frac{x_i - \bar{x}}{s_i^2} \sum_{j=1,\ j\neq i}^{n} w_{i,j}(x_i - \bar{x})$$

where x_i is the numerical attribute for spatial unit i (e.g., municipality-specific EM), $w_{i,j}$ is the assigned spatial weight between unit i and j, $\bar{x}$ is the global average of the attribute and s^2 is its standard deviation, that is:

$$s_i^2 = \frac{\sum_{j=1,\ j\neq i}^{n} (x_i - \bar{x})^2}{n-1}$$

where n is the number of spatial units under study. A standardized version (z-score) of I_i can be obtained by subtracting the mean of I_i from each individual autocorrelation and then dividing the difference by the standard deviation of I_i. As anticipated by these formulas, I_i investigates whether a unit of analysis is significantly different from its neighborhood. This statistic finds application in cluster-outlier analysis, where one wants to check for the presence of abnormal observations within a local spot of high spatial autocorrelation.

The expression of Getis–Ord G_i^* is:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{x}\sum_{j=1}^{n} w_{i,j}}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}}$$

which is already formulated as a z-score, so no further computation is required. G_i^* identifies statistically significant spatial clusters of high values (hot spots) and low values (cold spots) as compared to the global average. In this study, we preferred Getis–Ord G_i^* over local Moran's I_i, because our main goal was to detect the areas within or across Lombardy, Veneto and Emilia-Romagna that were hardest hit (hot) or relatively spared (cold) by the pandemic. However, we encourage the use of local Moran's I_i in ecological studies to assess the reasons for abnormal spatial trends in mortality or other health indicators.

As already mentioned, Getis–Ord G_i^* computes a z-score for each unit of observation. A high z-score, typically above 1.96 or 2.58, indicates a spatial clustering of high values; a low negative z-score, typically below −1.96 or −2.58, indicates a spatial clustering of low values. These thresholds are commonly adopted because 95% and 99% of the observations in a standard normal distribution lie between −1.96 and 1.96 and between −2.58 and 2.58, respectively. Values that exceed these boundaries are expected to be systematically above or below the overall average with a high level of confidence; on the contrary, a z-score ≈ 0 indicates no spatial clustering. That being said, a matrix W that expresses the spatial structure of geographic space needs to be determined to get these figures. Various types of spatial weight matrices have been proposed, and selecting the right one might be critical. In our analysis, we opted for the basic binary coding where a spatial unit is either a neighbor (1) or it is not (0), but more advanced features are possible in the R package spdep.

Another critical decision is the distance band (or threshold distance) that ensures that each spatial unit has at least one neighbor. Unfortunately, high bands may result in some units having too many neighbors, and this undermines the chances to detect local spots, while using low bands the largest and/or more isolated units can be excluded because they find no neighbors. We decided that a band of 12.5 km was a good compromise to depict a map pattern with adequate accuracy while excluding only 5 out of 2397 municipalities (Cortina d'Ampezzo, Chioggia, Porto Tolle, Comacchio and Casteldelci).

In order to perform the spatial correlation analysis, we used the R software (version 4.0.3). As listed in the R spatial_autocorrelation script provided in the supplement, the packages needed to perform the analysis are tidyverse, a collection of R packages designed for data science [11], haven, required for importing .dta files [12], spdep, required to carry out the spatial correlation [2], sp [9] and sf [8], required for managing shapefiles.

In the script, the first steps load the dataset containing the attribute of interest "estratio" and the shapefile dataset which includes coordinates of regions and municipalities. We used .dta files only because the EM estimation was performed using Stata, but our script may be easily modified to load other types of datasets as long as they share a common key with the spatial data and include the attribute of interest. In our case, the territorial unit identifier was the "municipality" variable and "estratio" was the attribute of interest. Please note that the Stata COVID19_EM script outputs "estratio" in single region datasets, while the R spatial_autocorrelation script loads a single dataset which can include the "estratio" of several regions. Thus, if the user wants to compute the spatial autocorrelation of several regions combined, their datasets obtained by the COVID19_EM.ado Stata procedure must be appended, either in Stata or in R, before running the spatial_autocorrelation script.

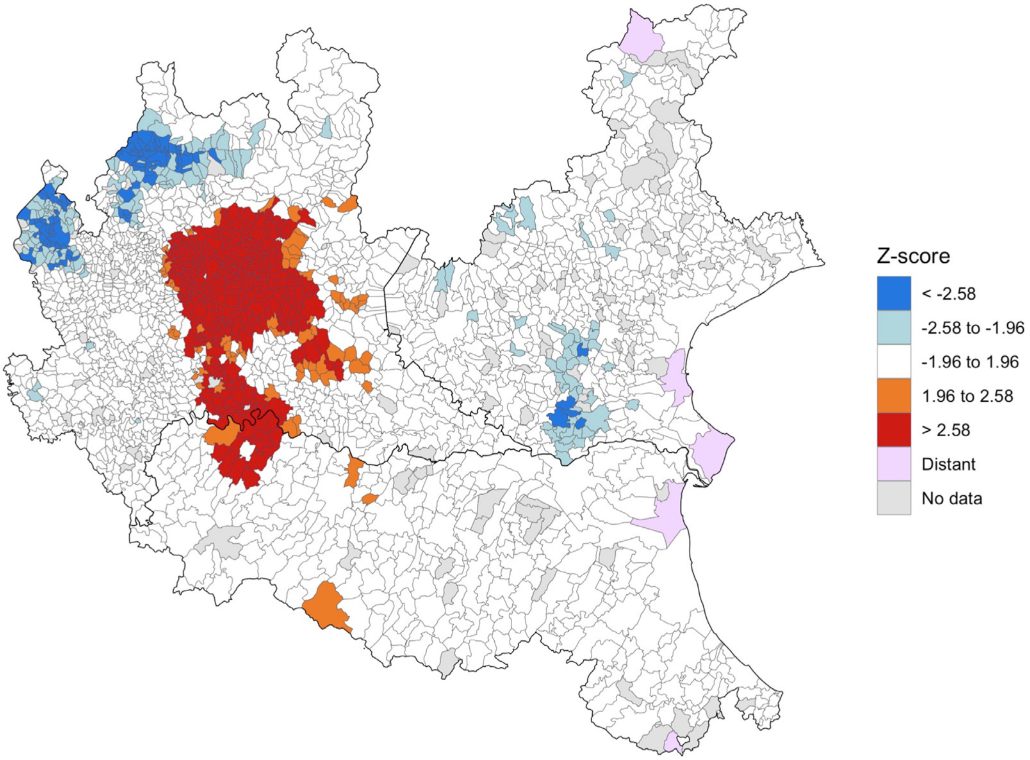


Fig. 2. Spatial autocorrelation of excess mortality of 2020 with respect to 2015-19 in the municipalities of Lombardia, Veneto and Emilia-Romagna, in the period February 23rd - April 30th and for males aged ≥ 75 years. Clusters of municipalities with low excess mortality (cold spots) are coloured in blue, those with high excess mortality (hot spots) are coloured in red.

To carry out the spatial correlation analysis two options of the spdep package need a customized setting: dnearneigh() and nb2listw(). The former creates the municipality centroids and joins neighbors in a distance range defined by the user specified values. We modified its default values using d1 = 0, d2 = 12,500, as previously declared. The nb2listw() option supplements a neighbors list with spatial weights chosen among a set of coding schemes. We modified this option using style = B (basic binary coding) and zero.policy = TRUE (allows the weights list to be formed with zero-length weights vectors).

Getis–Ord $G_i^*$was computed using the localG() function, also part of the spdep package. By this function the local autocorrelation is calculated for each municipality using the spatial weights object obtained as the output of nb2listw(). As a final step, the maps showing the spatial clusters are produced using the ggplot2 package, see for example Fig. 2 in which the spatial autocorrelation between municipalities of Lombardy, Veneto and Emilia-Romagna is displayed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2021.101257.

References

[1] L. Anselin, 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27, 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x.

[2] R.S. Bivand, D.W.S. Wong, Comparing implementations of global and local indicators of spatial association, Test 27 (2018) 716–748, doi:10.1007/s11749-018-0599-x.

[3] A. Getis, J.K. Ord, The analysis of spatial association by use of distance statistics, Geogr. Anal. 24 (1992) 189–206, doi:10.1111/j.1538-4632.1992.tb00261.x.

[4] ISTAT, 2020. Decessi e cause di morte: cosa produce l'Istat [online]. URL https://www.istat.it/it/archivio/240401 (accessed 10.29.20).

[5] P.A. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1950) 17–23, doi:10.1093/biomet/37.1-2.17.

[6] R. Newson, 1998-2021. PARMEST: Stata module to create new data set with one observation per parameter of most recent model. https://ideas.repec.org/c/boc/bocode/s352601.html.

[7] J.K. Ord, A. Getis, Local spatial autocorrelation statistics: distributional issues and an application, Geogr. Anal. 27 (1995) 286–306, doi:10.1111/j.1538-4632.1995.tb00912.x.

[8] E. Pebesma, 2018. Simple features for R: standardized support for spatial vector data. R J. 10, 439–446. doi:10.32614/rj-2018-009.

[9] E. Pebesma, R.S. Bivand, Classes and methods for spatial data in R, R News 5 (2005) 9–13.

[10] M. Pisati, 2008. SPMAP: Stata Module to Visualize Spatial Data. Stata Help Files. http://citec.repec.org/rss/bocbocodes456812.xml.

[11] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the Tidyverse, J. Open Source Softw. 4 (2019) 1686, doi:10.21105/joss.01686.

[12] H. Wickham, E. Miller, 2018. haven: Import and export SPSS, Stata, and SAS files. [online]. https://cran.r-project.org/web/packages/haven/index.html (accessed 11.16.20).