imajINE Integrative Mechanisms for Addressing Spatial Justice and Territorial Inequalities in Europe

Integrative Mechanisms for Addressing Spatial Justice and Territorial Inequalities in Europe

D2.2. Literature Review on Disaggregation Methodologies

Version 1

Authors: Esteban Fernandez-Vazquez, Maria Plotnikova, Paolo Postiglione, Fernando Rubiera-Morollon & Ana Viñuela

Grant Agreement No.: 726950

Programme call: H2020-SC6-REV-INEQUAL-2016-2017

Type of action: RIA – Research & Innovation Action

Project Start Date: 01-01-2017

Duration: 60 months

Deliverable Lead Beneficiary: UNIOVI

Dissemination Level: PU

Contact of responsible author: avinuela@uniovi.es

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 726950.

Dissemination level:
• PU = Public
• CO = Confidential, only for members of the consortium (including the Commission Services)

Change control

VERSION	DATE	AUTHOR	ORGANISATION	DESCRIPTION / COMMENTS

Bibliographic Information

Esteban Fernandez-Vazquez, Maria Plotnikova, Paolo Postiglione, Fernando Rubiera-Morollon, Ana Viñuela (2018). *Deliverable 2.2 Literature Review on Disaggregation Methodologies.* IMAJINE WP2 Analysis of Territorial Inequalities in Europe

This report can be downloaded from our website http://imajine-project.eu/

Acknowledgement

We thank the research assistants Alberto Cartone, Alberto Diaz-Dapena, Diana Gutierrez-Posada and Domenica Panzera.

December 2018

Acronyms and Abbreviations

AROPE	At risk of poverty or social exclusion
BIM	Bayesian Interpolation Method
CAR	Conditional Autoregressive (specification)
EU	European Union
EU-SILC	European Union Survey on Income and Living Conditions
GCE	Generalized Cross Entropy
GME	Generalized Maximum Entropy
MSE	Mean Square Errors
NUTS	Nomenclature of Territorial Units for Statistics
WP	Work Package

Table of Contents

Table of Figures & Tables

Introduction

Obtaining reliable estimates of data at fine spatial scales is essential for analytical and research purposes as well as for policy design and evaluation. Economic, demographic or education statistics are commonly available for the different types of *administrative units* or NUTS regions (Nomenclature of Units for Territorial Statistic) in which the European territory is divided. One of the main objectives of Work Package 2 (WP2) of the IMAJINE Project is to provide an inclusive and homogenous database at *local* level for several EU countries. This database includes information on two indicators that are essential for the study of territorial inequalities and which, until now, simply did not exist at local level for many EU countries: average household income and poverty. This report explains the methodology employed in the estimations, it underlines the novelties introduced and provides details of the procedure followed to create these variables at local scale.

The report is divided into four sections. Section 1 briefly reviews the literature on spatial disaggregation, distinguishing between techniques that simply try to obtain point estimates for areal units from those that produce estimates taking individual agents as units of analysis. Section 2 details the procedure that is applied in this WP2, where the information contained in a sample of households for several EU countries is projected onto the respective national population census and then adjusted to be consistent with the official regional aggregates. Section 3 shows the results obtained for four European countries (Portugal, Spain, United Kingdom and France), when the regional averages on household income and At Risk of Poverty and Exclusion (AROPE) indicators are disaggregated and estimated for *local* areas. The definitions of these local areas are different for each country, depending on data availability and their specific spatial configuration. Finally, Section 4 evaluates the reliability of the local estimates by comparing them to an existing 2011 local database for Spain and globally by conducting a Monte Carlo experiment. The results in both cases show how the proposed methodology for estimating data at local level provides statistically satisfactory socio-economic indicators.

1. A BRIEF REVIEW OF THE LITERATURE ON SPATIAL DISAGGREGATION OF DATA

Although essential in today´s economic analysis and policy evaluation, data at local level, i.e. beyond the NUTS regions, is restricted to few socioeconomic variables and only for specific years. Most of the variables are collected by the EU member states at a relatively broad or aggregated spatial level, mainly NUTs regions. The process by which information at a coarse spatial scale is translated to finer scales, maintaining the consistency with the original dataset, is known as spatial disaggregation.

In this document we will classify the techniques to disaggregate spatial data into two broad classes: the methodologies aimed at estimating the value for the local spatial unit of analysis (*areal techniques*) and the methodologies that take individual agents such as household and firmsas the unit of analysis (*micro-data based techniques*).

1.1 Areal techniques

Different areal interpolation techniques can be used in this context to transform data from a set of source zones to a set of target zones (see, e.g., Goodchild, M.F., Anselin, L. and Deichmann 1993; Goodchild, M.F. and Lam 1980). The first attempts to estimate economic variables at a disaggregated scale can be linked to the spatial smoothing methodology of Tobler (1979), which has been improved to include more complex structures (see King et al., 2004, for an exhaustive review).

Generally speaking, this type of methodology disaggregates a variable X of counts or totals across the A_i areas in each region i. To obtain these values it is necessary to obtain a function λ that varies across longitude (x) and latitude (y). The estimation criterion is to minimize equation 1:

$$\int\int \left[\left(\frac{\partial^2 \lambda}{\partial x^2}\right)^2 + \left(\frac{\partial^2 \lambda}{\partial y^2}\right)^2 \right] dxdy \quad (1)$$

Subject to the total amount of the variable X in each region.

$$\int_{A_i} \lambda(s) ds = X(A_i)\ ; \forall i \quad (2)$$

However, this type of procedure does not use any economic model to create the λ function, as it is just a graphical distribution of the totals over a map. Several methodologies have tried to obtain disaggregated estimates reducing inaccuracy problems, and most of these can be divided into direct and indirect estimates (see Rao 2003). In the direct estimate methodologies there are two options, the 'model based' estimators and the 'design based' estimators (see Pfeffermann, 2013, for a complete summary of the most recent developments in these two methodologies). Model-based estimators (as in Royall, 1970) try to extrapolate weights of each sub-area using an econometric model with other support variables that are related to the weight of the area. However, as shown in different articles (e.g., Hansen et al. 1983) these estimators face the important risk of high bias if the model is misspecified.

Due to the risks involved with model-based estimates, one of the best-known estimators is the design-based methodology. This is a statistical approach that tries to obtain optimum sample weights in order to implement a sampling design. According to this methodology, small area estimations can be obtained when each sample is assigned an unbiased weighting. Some recent proposals within this methodology can be found in Jiang and Lahiri (2006) or Chandra and Chambers (2009), while a detailed review can be found in Rao (2003).

The main problem of direct estimators is that they usually result in wide confidence intervals due to problems of small sample size (even in the case of a correctly-estimated model). They also assume that it is possible to modify the design of the sampling process, and it is not obvious how to choose the weights. Hence, it was necessary to devise methodologies that reduce these problems. The indirect method incorporates previously-estimated information with out-of-the-sample data to adjust the estimations, thereby reducing the problems of variability in the estimations. A good example of this methodology can be found in Griffiths (1996).

A few areal interpolation techniques consider the special features of spatial data. Specifically, the spatial dependence effect could provide useful information in the spatial disaggregation procedure. Spatial dependence reflects a situation where values observed at one location depend on the observations at nearby locations. Benedetti and Palma (1994) introduced the Bayesian Interpolation Method (BIM) to the areal interpolation problem which exploits this general property of spatial data. For recently proposed areal interpolation methods that consider spatial nature of data see Gotway et al. (2013) and Murakami and Tsutsumi (2011).

BIM requires some assumptions on the spatial data generating process. Commonly, spatially-referenced data are considered to be a realization from a spatial stochastic process or random field, which is a collection of random variables indexed by their locations. When dealing with the areal interpolation problem, data related to both source and target zones can be interpreted as realizations of spatial stochastic processes. The spatial stochastic process generating the data related to the target zones (i.e., the areal units corresponding to the finer spatial scale) is referred to as the *original process*. The spatial stochastic process generating the data for the source zones (i.e. the areal units corresponding to the aggregated spatial level) is referred to as the *aggregated process*. Assuming that data are available only at the aggregated spatial level, the objective becomes to restore the realizations of the original process given the realization of the aggregated one.

The basic assumption on which BIM relies concerns the joint probability distribution of the original process, which is assumed to be a Gaussian distribution. The spatial dependence effect is taken into account by modelling the Gaussian random field by the Conditional Autoregressive (CAR) specification. This assumption does not entail any loss of generality since any Gaussian process on a finite set of sites can be modelled according to this specification. The CAR specification introduces the spatial dependence effect in the covariance structure of the process as a function of a scalar parameter of spatial autocorrelation and of a spatial weight matrix**,** which summarizes the proximity between any pairs of spatial units. Following a Bayesian approach, the prior information on the distribution of the original process is combined with the data available at the aggregated spatial level to derive the posterior probability distribution of the original process. Benedetti and Palma (1994) derive the parameters of this posterior

probability distribution, which are the BIM estimates. Any inference on the original process can be based upon the specified posterior distribution.

One problem with the class of techniques detailed above is that they do not allow for a rich analysis of the spatially disaggregated data: they produce estimates of the variable of interest at the desired spatial scale, but use a large scale of analysis to study its heterogeneity between different groups of individuals or to quantify inequalities within these small areas. In order to overcome this problem, the type of spatial disaggregation techniques applied in this WP2 follow an alternative approach that exploits the availability of microdata.

1.2 Micro-data based techniques

The spirit of this second family of disaggregation techniques is different. Instead of producing direct estimates of the variable of interest at the desired spatial scale, the basic idea is to predict this variable at the level of individual agents (such as households, firms, workers, etc.). If the geographical location of these individual agents is observable at a highly disaggregated spatial scale, the indicators for the small areas are calculated simply by summing or averaging the individual estimates. In most cases, however, researchers have to deal with microdata that does not provide the location of the individuals at a high level of detail. Surveys designed to study income distribution issues (such as household surveys) do not usually allow for a precise geographical location of the individuals surveyed. On the other hand, databases that do allow for a more precise spatial location of the individuals, such as the microdata of a Population Census, do not normally contain information on economic variables like household income. The most important works that proposed a way to solve this problem are the contributions by Elbers et al. (2003) and the modification proposed later in Tarozzi and Deaton (2009). The basic idea of both works consists of "projecting" predictions of the variable of interest for a household survey onto the sample of households that form the population. In a nutshell, the procedure consists of three steps:

1. From the household surveys (HS), estimate a model of your variable y of interest, y=(X), where X is a set of regressors observable in the Population Census (PC).
2. Recover the set of parameters β estimated on the HS (with some degree of heterogeneity across regions or clusters of households) and take them to the PC.
3. Given the X observable in the PC and the corresponding $\hat{\beta}$, predict the figures of y for the households surveyed in the PC ($\hat{y} = X\hat{\beta}$).

The estimates produced have the advantage of greater precision than previous methodologies due to the large number of households in the census (see Tarozzi and Deaton, 2009). Additionally, this large number of estimates permits the study of potential differences between the individuals or household belonging to the same small area.

This feature is highly appealing for the social researcher, and this is the approach that we follow in this WP. A problem that arises with applying this type of estimation, however, is that the estimates may not even be consistent with the aggregates that are already observable: once the techniques produce estimates at household level for a given region, for example, the mean value

of these estimates are not necessarily equal to the mean household income available in the official databases. This implies that the estimates produced by applying this procedure are not a true disaggregation of the spatial data.

In order to overcome this limitation, we propose the use of an additional step to the procedure depicted above to adjust the estimates to official observable aggregates. This adjustment allows us to incorporate information from the observable aggregates in order to make the *n* estimates consistent with it. We propose to make this correction through a Generalized Cross Entropy (GCE) estimator based on Bernadini-Papalia and Fernández-Vázquez (2018), which is detailed in the next section.

2. SPATIAL DISAGGREGATION ESTIMATIONS OF DATA USING GENERALIZED CROSS ENTROPY (GCE) METHOD

This methodology is based on the framework of Maximum Entropy. In this framework, the variable of interest has a probability distribution with an unknown probability for each value. The basic idea of this type of methodology is to obtain the estimation with the highest degree of uncertainty that at the same time is able to fulfil the conditions from observable data. So, the set of probabilities has to be calculated through optimization of an Entropy function as in equation (3) (see Shannon, 1948 for additional details).

$$Ent(p) = -\sum_{m=1}^{M} p_m \ln(p_m) \tag{3}$$

This function has its maximum value for a distribution of probabilities p_m of a discrete random variable that distributes uniformly with M different possible values. Hence, the optimization process would tend to divert the minimum distance possible that the restrictions allow. Any new information about the variable of interest would restrict the feasible region of the optimization problem, moving the optimum value away from the uniform distribution.

With this same idea, the Generalized Maximum Entropy (GME) estimator (see, Golan et al., 1996) has been applied to the estimation of linear regression equations. It considers the coefficients of a linear regression β_k as a discrete random variable with different M possible values for each k coefficient. The optimization problem would recover the probability of the different possible values of the regression coefficients. Once the vector of probabilities for each coefficient p_{km} has been calculated, the estimated set of coefficients $\tilde{\beta}_k$ can be calculated as its expected value given the estimated probabilities for each value of their discrete distribution:

$$\tilde{\beta}_k = \sum_{m=1}^{M} p_{km} b_{km} \; ; \; k = 1, \dots, K \tag{4}$$

GME redesigns the linear regression estimation as an optimization problem of the entropy function in equation (3) subject to the linear relationship between the K exogenous (x_k) and the dependent (y) variables in a sample of $i = 1, \dots, N$ datapoints. This methodology does not need restrictive assumptions for the error, and just requires a finite matrix of variance and covariances as well as a null expected value in the errors. With these assumptions, the errors are also presented as a discrete random variable with J possible values with a set of probabilities u_j for possible values of the error presented in v_j. The final optimization problem in GME for a cross-section database is summarized in equations (5), (6), (7) and (8).

$$\underset{\boldsymbol{P},\boldsymbol{U}}{Max}\, Ent(\boldsymbol{P},\boldsymbol{U}) = -\sum_{k=1}^{K}\sum_{m=1}^{M} p_{km}\, ln(p_{km}) - \sum_{i=1}^{N}\sum_{j=1}^{J} u_{ji}\, ln(u_{ji}) \quad (5)$$

subject to:

$$y_i = \sum_{k=1}^{K}\sum_{m=1}^{M} b_{km} p_{km}\, x_{ki} + \sum_{j=1}^{J} v_j\, u_{ji}\ ;\ i = 1, \dots, N \quad (6)$$

$$\sum_{m=1}^{M} p_{km} = 1\ ;\ k = 1, \dots, K \quad (7)$$

$$\sum_{j=1}^{J} u_{ji} = 1\ ;\ i = 1, \dots, N \quad (8)$$

Through this optimization problem, GME finds an optimal solution that is compatible with the feasible region defined by the linear relationship between the observed values in the dependent and the exogenous variables.

We apply a modification of the GME procedure defined in Bernadini-Papalia and Fernández-Vázquez (2018) as a consistent method to update the estimates produced after applying the three-step process described in Elbers et al (2003) and Tarozzi and Deaton (2009). The variable or indicator of interest y is estimated for a set of N individual data points (households, for example), which once aggregated or averaged are not consistent with an observable aggregate $\bar{y}$. Let us denote these estimates as $\hat{y}_i$. The GME estimator assumes that estimates y_i are just a realization of a wider set of possible results. This set of possible realizations is contained in a support vector $\boldsymbol{b_i}$, centered symmetrically around $\hat{y}_i$. Note that since $\hat{y}_i$ has been obtained applying some type of regression analysis, (like Ordinary Least Squares, for example), it is relatively easy to define natural values for the elements on this vector b_i. For the sake of simplicity, but without loss of generality, let us consider that $M = 3$ values are included in $\boldsymbol{b_i}$,with y_i being the central one and the limits of this vector defined by the expression $\hat{y}_i \pm 3\sigma_y$, where σ_y denotes the standard deviation calculated for the prediction $\hat{y}_i$.

This is a way of including some flexibility in the point-estimates: by assuming that within the range of three standard deviations of our predictions other possible estimates could have been obtained, we incorporate some natural uncertainty to our point estimates that will allow the logic of the GME procedure to be applied. The idea is to assume that each realization on this vector (b_{im}) has some probability of occurring (p_{im}) and that this probability can be estimated by applying some optimization criterion. Once the probabilities are estimated, the value of the variable of interest for the data point i is defined as:

$$\tilde{y}_i = \sum_{m=1}^{M} p_{im} b_{im} \quad (9)$$

In particular, the GME estimator will choose the distribution of probabilities that least deviates from a situation of maximum uncertainty as the optimal solution: assuming that the reference is a uniform distribution where all the M values in the vector $\boldsymbol{b_i}$ are equally probable, the GME estimator will depart from this solution only if some additional information forces it to do so. This additional information can come in the form of some aggregate that should be consistent with the $\tilde{y}_i$ estimates. If no additional information is included, note that the GME produces the same solution as the initial estimates $\hat{y}_i$.

Going to the details of this methodology, the estimates for the $i = 1, \dots, N$ households in the sample can be expressed as in equation (9) based on a support vector defined following the logic depicted above. The GME estimator defines a feasible region where the mean (weighted by sampling factors N_i) of the N observations has to be equal to an out-of-sample aggregate $\bar{y}$. As in GME, this model includes an error with similar assumptions, but the notation is adapted to the problem. However, the restriction has been modified, given that the weighted mean of the variable of interest has to be equal to the observable aggregate.

$$\underset{\boldsymbol{P}}{Max}\, Ent(\boldsymbol{P}) = -\sum_{i=1}^{N}\sum_{m=1}^{M} p_{im}\, ln(p_{im}) \quad (10)$$

subject to:

$$\bar{y} = \sum_{i=1}^{N}\left[\sum_{m=1}^{M} p_{im} b_{im}\right]\left[{}^{N_i}/_{N}\right] \quad (11)$$

$$\sum_{m=1}^{M} p_{im} = 1;\ i = 1, \dots, N \quad (12)$$

Note that the situation described in general terms above is the type of problem that we find when the methodology of Elbers et al. (2003) or Tarozzi and Deaton (2009) is applied to disaggregate some socio-economic indicators produced for the European Union (EU). The survey on issues related to income distribution, poverty or, more generally, living conditions, is the Statistics on Income, Social Inclusion and Living Conditions (EU-SILC).[i] This survey is rich on information about several characteristics of individuals and the households and it is the most commonly-used database when questions regarding inequality are studied. However, due to confidentiality issues, the EU-SILC only provides aggregated information about the geographical location of the individuals surveyed: depending on the specific country, information about the location is only released at the level of NUTS1 or NUTS 2 regions. Since this scale could be considered as insufficient to study spatial inequalities across the EU, a GME estimation designed

to produce spatially-disaggregated indicators is applied. The set of indicators produce will be consistent with the mean regional values $\bar{y}$ reported in the EU-SILC (at the NUTS1 or NUTS2 levels). Details of the data and the procedure followed are explained in the next section.

3. AN APPLICATION: SPATIAL INEQUALITIES IN UNITED KINGDOM, FRANCE, SPAIN AND PORTUGAL

3.1 Data

The empirical illustration is carried out by combining two databases for 2011, the EU-SILC provided by Eurostat and the microdata contained in the Population Census of four different countries (United Kingdom, France, Spain and Portugal), which are provided by their respective national statistics institutes.[ii]

This research focuses on two variables of interest: household income and the propensity of being *at risk of poverty and exclusion* (hereinafter AROPE) in 2011. Regarding this last indicator and following the definition given by Eurostat, for a household to be considered AROPE, one of the following three conditions must hold:

1. a disposable income below 60% of the national median,
2. being severely materially deprived, or,
3. living in a household with a low work intensity.

As commented above, the main problem is that the observations in the microdata of the EU-SILC can only be located at a NUTS1 or NUTS2 level, depending on the country, which prevents the study of spatial differences in the two indicators of interest at a more detailed spatial scale.

The procedure explained in the previous section is applied in order to spatially disaggregate these indicators. Census information provides two advantages that make them suitable for this procedure. The first one is that it possesses the necessary information about the location of the household, while the second is that it produces a set of household estimates coming from a database characterized by a large sample size. Table 3.1 summarizes the sample sizes in both EU-SILC and census in the different countries of the analysis.

Table 3.1: EU-SILC and census sample size and location level – 2011.

	Spain	United Kingdom	France	Portugal
Household survey				
Households in SILC	13,109	8,058	11,360	5,740
SILC-Location level	NUTS 2	NUTS 1*	NUTS 2	Country
Number of locations	19	37	12	1
Census microdata				
Households in census	1,619,806	1,312,291	8,741,050	204,409
Census location level	Municipalities	Local areas	Canton	NUTS 3 + 5 cities
Number of locations	588	296	3708	37
Source	Spanish National Statistical Institute - INE	UK National Institute - ONS	National Institute - INSEE	IPUMS

*Note: Due to aggregation of local entities in the census.

Although census data generally provides information with a high degree of disaggregation, the spatial unit in each census is not necessarily the same across countries. The spatial unit of analysis depends on both the respective national administrative division and the information available about the place of residence. As a result, the disaggregation is made for the municipalities of Spain, local areas of the UK, *cantons* of France and NUTS3 and the five most important cities of Portugal.

The variables used to disaggregate the variables of interest are similar to the ones in Tarozzi and Deaton (2009), with an effort made to include all the possible relevant variables in the census that also appear in the EU-SILC with an identical definition. In addition, the censuses may differ in terms of available information and definitions. Thus, in order to avoid a conglomerate of heterogeneous processes of estimation, the variables have been chosen with an eye to having similar variables - or at least concepts - in all the countries where possible. This effort makes the methodology more consistent between countries. However, there are cases where some information was discarded because no other country provided any similar concept.

The set of exogenous variables can be divided into two groups, where the first contains characteristics of the head of household and the second comprises characteristics of the household as a whole. The household head is defined, in order, by status in employment,[iii] hierarchy, level of education, age and gender. UK national institutes provide their own identification of the household reference person according to labour status. In this case, the EU-SILC identification of the household head follows the same criteria.

The chosen variables for each country are reported in Table 3.2. It can be seen that despite the differences between the national censuses, the set of variables is virtually the same in all cases: labour status, personal characteristics of the head of household, and structure of the household. However, some minor difference can be found in each case. To accommodate these differences, the GME estimation with the EU-SILC in each country is carried out with the available information of its corresponding census.

Table 3.2: Variables used as predictors – EU-SILC and census microdata (2011)

Head age (and its square)
Head gender
Head is foreigner from an EU/ non-EU country
Head marital Status: Married/ Separated/ Widow/ divorced
Head education: Post-mandatory non-college education and college education
Head activity status: Worker, Retired or disable, other activity
Head is in part time employment
Head occupation: Manager, Technician or professional, Support worker or sales, Craft, machine operators or skilled agricultural worker
Head economic sector - CNAE (1 digit)
House tenancy status
Number of rooms in the house
Number of workers in the household
Number of members (by age) in the household
Household structure: single parent/ couple with/without children and other with/without family

Note 1. Sectors are defined according to the statistical classification of economic activities in the European Community (NACE).

Note 2. Due to census definitions, immigrant variables and divorced/separated dummies have been grouped for France. Occupation in France were not used due to differences between census and EU-SILC.

Note 3. Household structure and tenure has been reduced when the national census does not provide information for all the dummies.

3.2 Results

As indicated above, the level of disaggregation in the EU-SILC depends on the information that each country provides to Eurostat. Depending on the specific case, it is provided at NUTS1, NUTS2 or just country level. As an illustration, the original aggregates from EU-SILC are provided in Figures Figure 3.1 andFigure 3.2.

Figure 3.1: Mean Household Income – Portugal, Spain, France and United Kingdom, 2011. Regional aggregates.

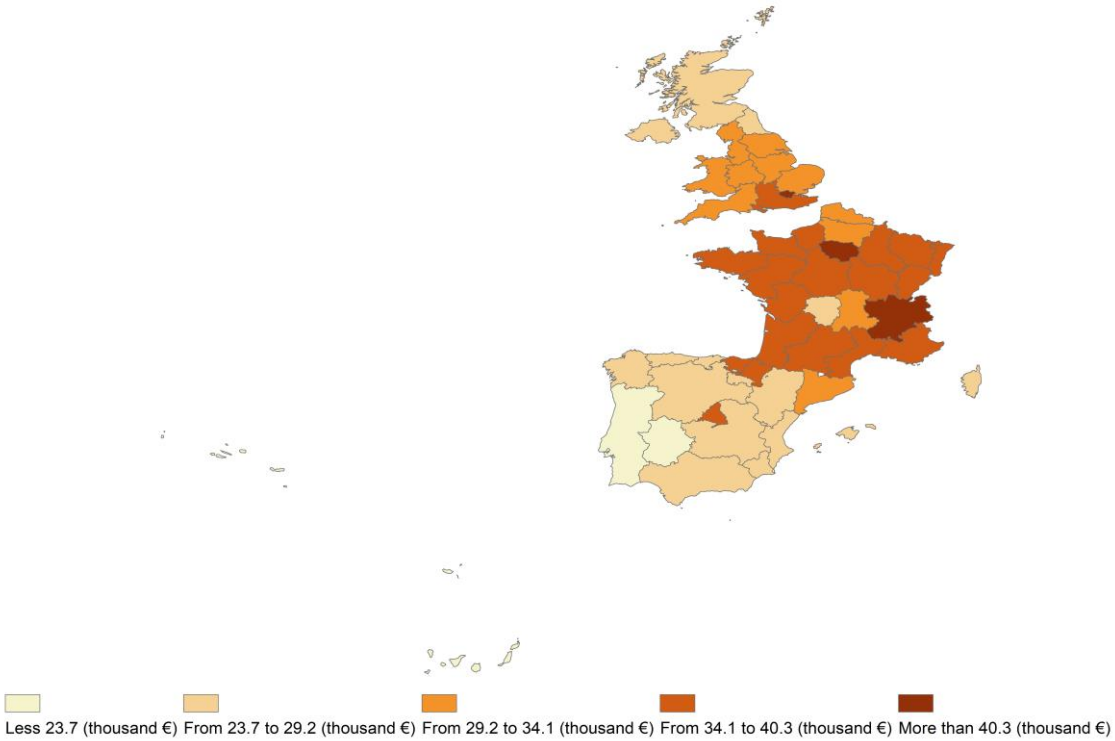
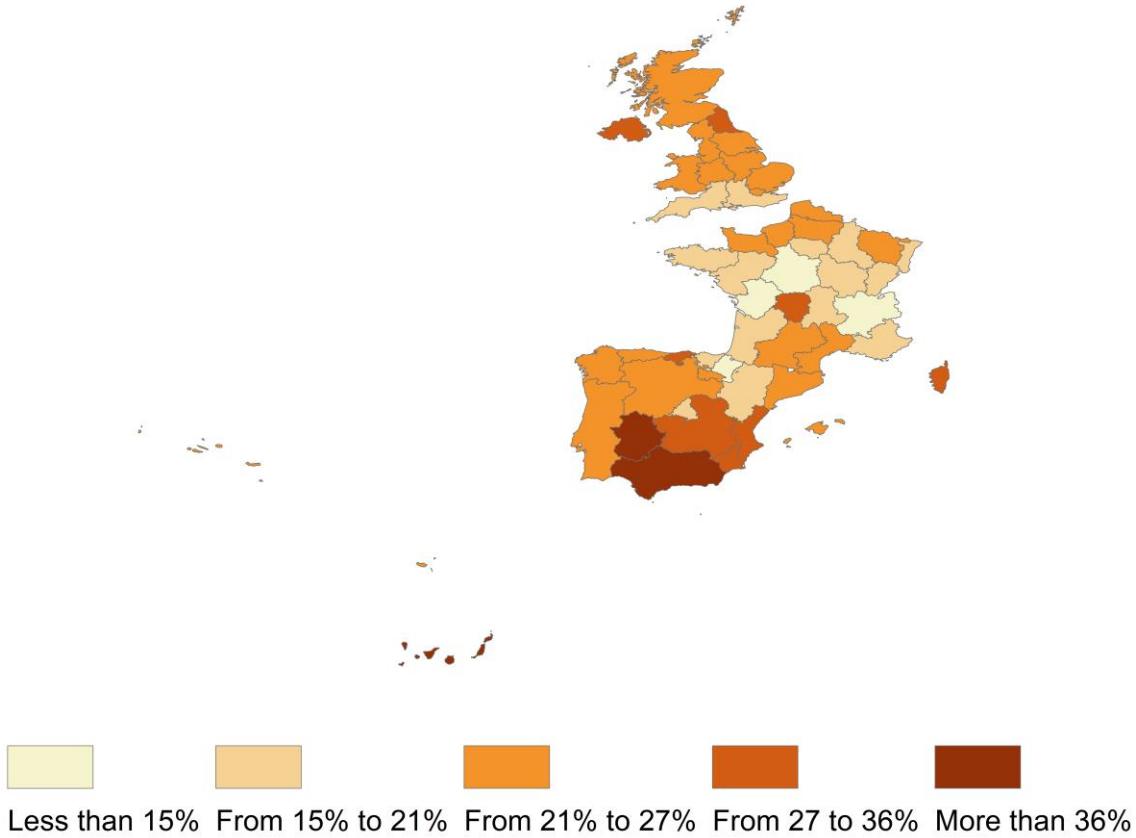


Figure 3.2: AROPE – Portugal, Spain, France and United Kingdom, 2011. Regional aggregates.



In order to perform the disaggregation of the data, we should first have some indication of an important degree of variability in the data within the EU-SILC regions. This variability would point to a possible non-stochastic distribution of the data within these regions. Table 3.3 indicates the percentage of variation for household income in the EU-SILC that can be found within the regions.

Table 3.3: Theil decomposition of the income.

Indicator	Spain	France	Portugal	United Kingdom
Theil	0.241	0.248	0.256	0.274
Theil between	0.010	0.007	Not applicable	0.016
Theil within	0.231	0.242	0.256	0.258
Proportion within	95.82%	97.28%	100%	94.23%

This table decomposes the variability in household income of in the EU-SILC *between* and *within* the identifiable regions of this database. In the case of Portugal, it is not possible to identify any region, so all the variability is within groups. According to this decomposition, it can be seen that an important part of the variability (more than 90%) can be found within the regions identified in the EU-SILC. Of course, this result does not indicate that the spatial distribution of the income

within the regions is homogeneous. However, it does indicates that there is plenty of remaining information that may be explained in part by an additional spatial level in the analysis.

By applying the GME estimation procedure, regional aggregates are disaggregated to the same spatial level that can be found in the respective national population census. These disaggregated indicators are plotted in the Figure 3.3 and Figure 3.4.

Figure 3.3: GME estimation of Mean Household Income – Portugal, Spain, France and United Kingdom, 2011. Spatial disaggregation.

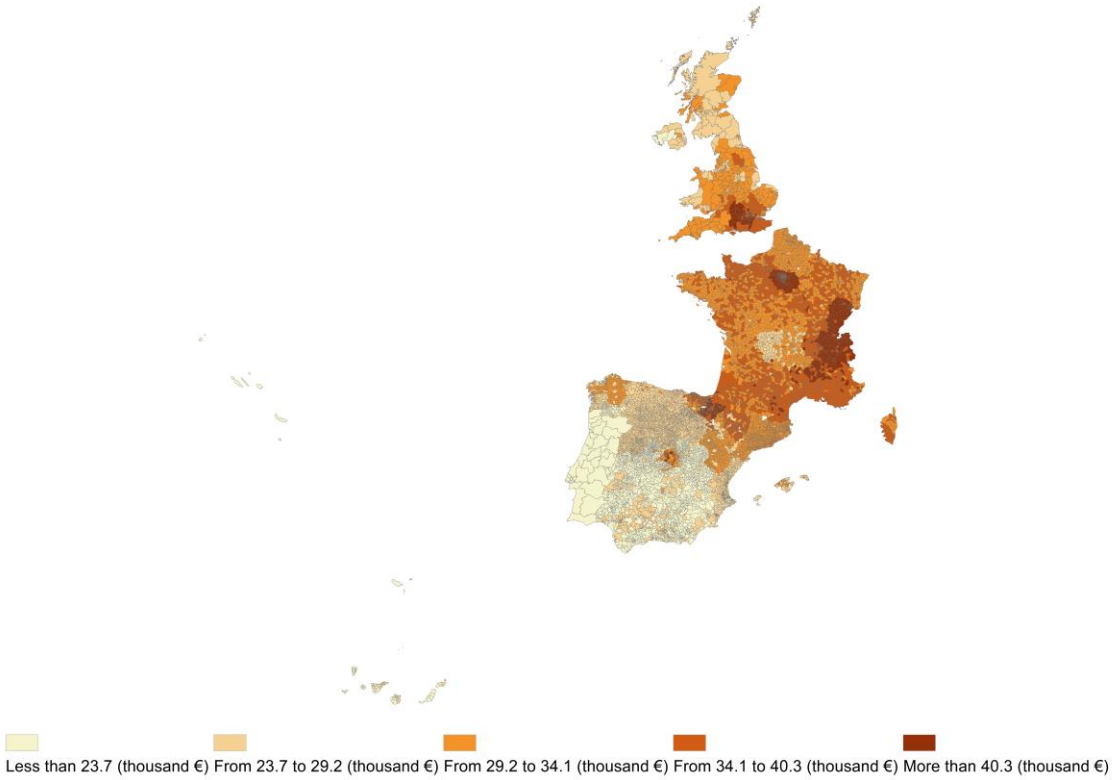


Figure 3.4: GME estimation of AROPE – Portugal, Spain, France and United Kingdom, 2011. Spatial disaggregation.

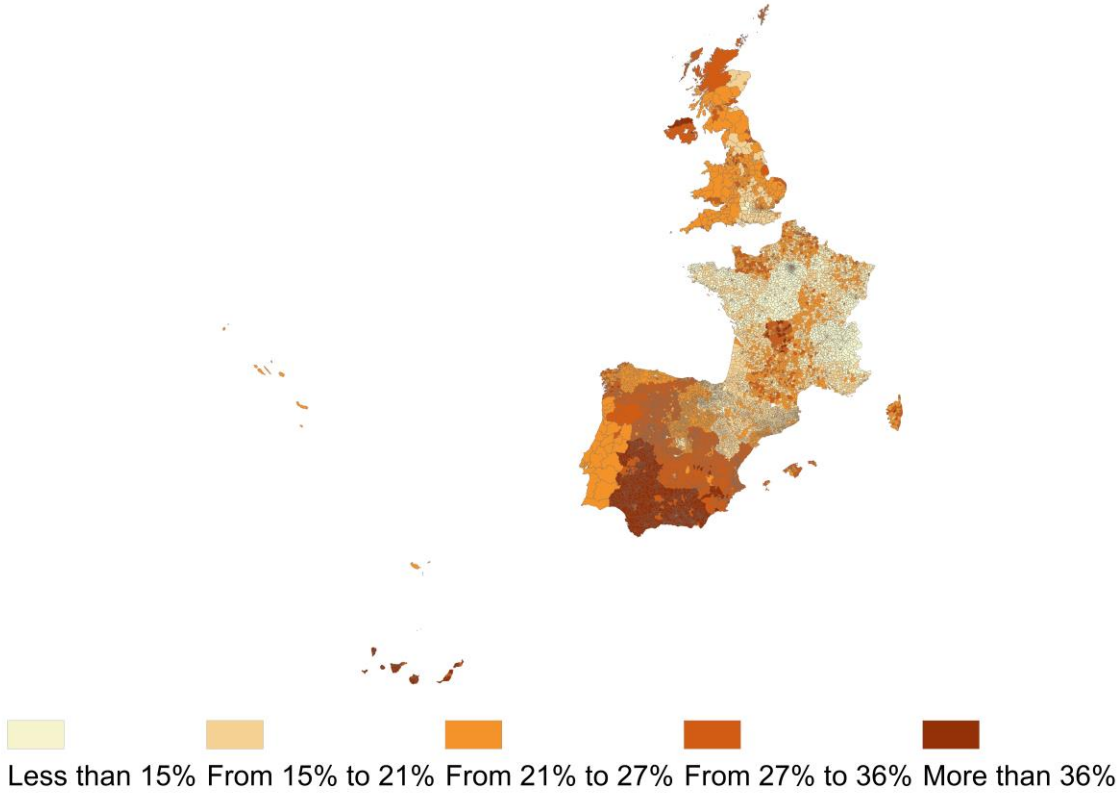


Figure 3.3 and Figure 3.4 respectively represent the spatial distribution of household income and AROPE at local level. This type of analysis allows for a more detailed investigation of spatial differences than the regional aggregates presented in the official statistical databases usually allow. This is especially important in countries where the spatial configuration is defined by administrative regions with high internal heterogeneity.

As an example, take the case of Andalusia in Spain: this NUTS2 region had more than 8 million inhabitants in 2011, and having only a regional mean of household income or a regional average AROPE rate provides a quite simplistic summary of the situation of the region. Disaggregating these regional aggregates allow us to identify the potential idiosyncratic patterns of some specific sub-regions within the NUTS2 division (e.g., rural areas versus urban agglomeration, differences between coastal areas heavily dependent on tourism and the rest of the region, etc.).

4. EVALUATION: A CASE OF STUDY AND NUMERICAL SIMULATIONS

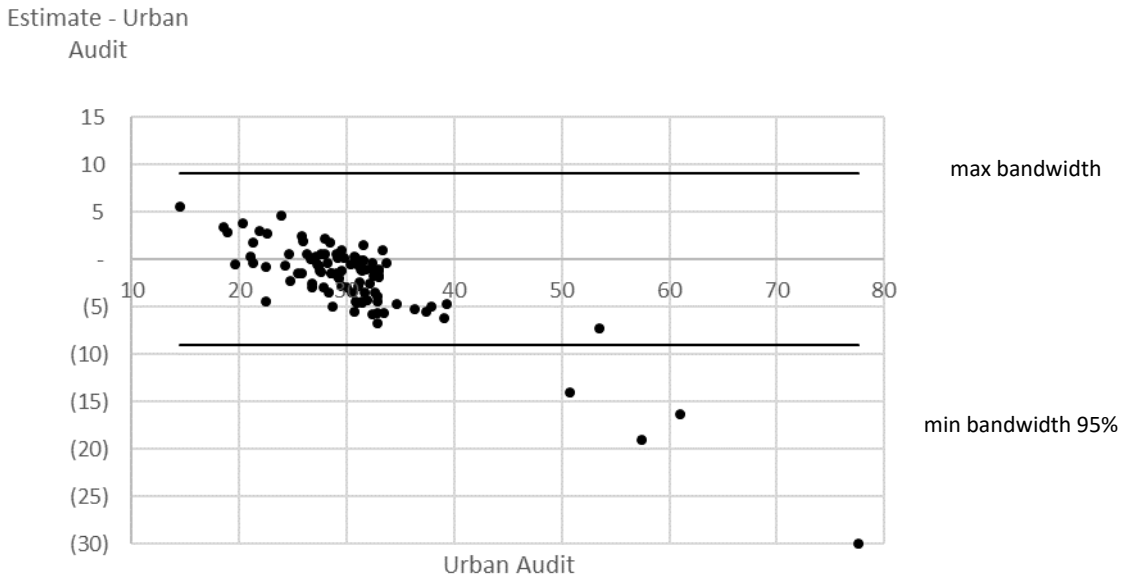
The reliability of the estimated data will be evaluated in this section. The idea is to quantify the potential error arising from the application of the estimation procedure depicted in Section 2 and implemented in Section 3. There are two possible sources of error: a poor specification of the econometric model applied to the household survey, and any significant differences between the households of the census and the household survey. The task of measuring this error is somewhat problematic since there are no "true" observable data to compare our estimates with (otherwise, of course, the estimation would not be necessary). We tackle this situation by following two different strategies: selecting cases studies where some data are observable, and conducting numerical experiments. The following sub-sections develop these two strategies.

4.1 Household income in Spain 2011: the Urban Audit database for large municipalities

Despite the previously mentioned lack of data on income at local level, there are a number of cases in which is possible to obtain them for some specific locations. As a case study, we have taken the estimates of household income at municipal level contained in the Urban Audit database in Spain (from INE), which provides information on several socio-economic indicators for a set of 126 municipalities, corresponding to those with more than 20,000 inhabitants.[iv] This information can be used to check the accuracy of our estimations for this group of municipalities.

The purpose of this comparison is simple: we take the figures contained in the Urban Audit database as the "true" values of the mean household income for this group of Spanish cities and we compare them with our estimates for the same group of municipalities. All these comparisons refer to 2011. Although there are differences in the way of estimating these figures (our estimates disaggregate the regional figures of household income reported in the EU-SILC, while the data in Urban Audit basically uses fiscal data), one would expect to obtain similar values. Figure 4.1 represents the gap between our estimates and the reported values of the mean household income for each municipality analysed in the Urban Audit database. On the vertical axis we plot the gross difference between these two figures for each one of the 126 municipalities that are plotted in the horizontal axis. Ideally, all the points represented in this graph should be on the horizontal line that crosses the vertical axis at zero. More practically, having differences bounded within some sensible limits would be an indication of having "good estimates".

Figure 4.1: Difference between urban audit (x) and estimates with EU-SILC (y) for household income (thousand €)



The maximum and minimum bandwidths for this comparison have been calculated, in accordance with Demombynes et al. (2007), as two times the standard deviation of the sample mean for the values reported in the Urban Audit database. Assuming a Gaussian distribution, this should bound 95% of the values coming from the same data generation process. An indication of similarity between our estimates and these reported values would be that our estimates lie within these limits. The errors reported in Figure 4.1 do not indicate any important bias in the estimations, and only 4 out of the 126 observations has a value outside of the bandwidth. These comparatively large errors correspond to four of the richest municipalities of Spain, and a possible reason for the bias is the location effects in these cities which cannot be represented with personal information, regional heterogeneity or the constraints.

4.2 Numerical experiments

In addition to the empirical evidence about the reliability of the methodology presented in the previous subsection, some more general evidence about the performance of the proposed methodology is desirable. This subsection presents a Monte Carlo experiment that provides a more general evaluation of the procedure. For the sake of simplicity, and also in order to make a straightforward comparison with the original formulation in Tarozzi and Deaton (2009), we replicate the general characteristics of the Monte Carlo simulations conducted in that paper (see Tarozzi and Deaton, 2009, pp. 781-784) .

The idea of this simulation is to generate "true" but not completely observable data for a set of households and small spatial areas. The population simulated consists of $N = 10{,}000$ households distributed uniformly across 100 small (local) areas. The values of the variable of interest (y) is generated as:

$$y_{hc} = 20 + x_{ch} + e_c + \varepsilon_{ch} \quad (13)$$

where the subscript c stands for the local areas and h for the households. e_c is an area-specific shifter that distributes as $e_c \sim N(0,0.01)$ and ε_{ch} is a disturbance that distributes normally with mean zero and variance $\sigma^2(x)$, where:

$$\sigma^2(x) = \frac{exp\left[0.5x_{ch} - 0.01x_{ch}^2\right]}{1 + exp[0.5x_{ch} - 0.01x_{ch}^2]} \quad (14)$$

The predictor x_{ch} is generated as $x_{ch} = 5 + z_c w_{ch} + g_c$, where $z_c, g_c \sim U(0,1)$ and $z_c \perp g_c$ and $w_{ch} \sim N(0,1)$.

The partial information that we have about this data generation process is a random sample of households, assuming that we sample 10 household on every local area (this makes a sample size of $n = 1{,}000$ in total). Note that this sample plays the role of the household survey in the methodology of Elbers et al. (2003) and Tarozzi and Deaton (2009). On this sample we run regression equations that estimate the parameters in (13), to later use these estimates on the whole population of $N = 10{,}000$ households. Note that this population plays the role of the census in the methodology, where the geographical location of the household (each local area c) is observable. Once the predictions for the whole population are calculated, this allows estimates to be obtained of some average for the variable of interest on each c.

In order to evaluate their method, Tarozzi and Deaton (2009) compare the differences between the true values of y_{hc} and the predictions obtained by applying their methodology in comparative terms to Elbers et al. (2003), finding that it produced smaller bias and Mean Square Errors (MSE).

Following this idea, we have extended this numerical experiment by including the proposed GME procedure into the comparison. In the Monte Carlo experiment we apply the methodology explained in Section 2, assuming that the mean value of y ($\bar{y}$) for the population is observable. We apply the GME program depicted in equations (10)-(12) to each simulation drawn. As in Tarozzi and Deaton (2009), we take 250 replication in the experiment, and a summary of the results is reported in Table 4.1:

Table 4.1: Monte Carlo simulation, 250 replications

Method	Bias	RMSE
Tarozzi and Deaton (2009)	-0.001	0.127
GME	0.000	0.087

Note. Bias is defined as the mean difference along the 250 replications of the true and estimated values of the variable of interest. RMSE stands for the square root of the mean square error along the 250 replications. See Tarozzi and Deaton (2009), page 782, for further details

The results of the numerical simulation suggests that the applied GME adjustment manages to improve the estimates produced by the methodology developed in Tarozzi and Deaton (2009), both in terms of bias (although only marginally) and RMSE. This results is not surprising given the general properties of the GME estimators and the fact that it exploits some piece of additional information, namely the aggregates ($\bar{y}$) that the estimates should be consistent with.

REFERENCES

Benedetti, R., & Palma, D. (1994). Markov random field-based image subsampling method summary. Journal of Applied Statistics, 21(5), 495–509. doi:10.1037/a0017771

Bernadini-Papalia, R., & Fernández-Vázquez, E. (2018). Information theoretic methods in small domain estimation. Econometric Reviews, 37(4), 347–359.

Chandra, H., & Chambers, R. (2009). Multipurpose Small Area Estimation. Journal of Official Statistics, 25, 379–395.

Demombynes, G., Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2007). How Good a Map? Putting Small Area Estimation to the Test (No. WPS4155). World Bank Working papers series. Washington D.C.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. Econometrica, 71(1), 355–364.

Golan, A., Judge, G., & Miller, D. (1996). A maximum Entropy Econometrics: Robust Estimation with limited data. New York: Wiley.

Goodchild, M.F., Anselin, L. & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. Environment and Planning A, 25, 383–397.

Goodchild, M.F. and Lam, N. S. (1980). Areal interpolation: a variant of the traditional spatial problem. Geo-Processing, 1, 297–312.

Gotway, C. a, Young, L. J., Journal, S., Statistics, G., Mar, N., & Young, J. (2013). A Geostatistical Approach to Linking Geographically Aggregated Data from Different Sources. Journal of Computational and Graphical Statistics, 16(1), 115–135.

Griffiths, R. (1996). Current population survey small area estimation for congressional districts. In Proceedings of Section on Survey Research Methods, American Statistical Association (pp. 314–319).

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association, 78(384), 776–793. doi:10.1080/01621459.1983.10477018

Jiang, J., & Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. Journal of the American Statistical Association, 101(473), 301–311. doi:10.1198/016214505000000790

King, G., Rosen, O., & Tanner, M. A. (2004). Ecological Inference: New Methodological Strategies. Analytical Methods for Social Research. Cambridge University Press. doi:10.1017/CBO9780511510595

Murakami, D., & Tsutsumi, M. (2011). A new areal interpolation method based on spatial statistics. Procedia - Social and Behavioral Sciences, 21, 230–239. doi:10.1016/j.sbspro.2011.07.034

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. Statistical Science, 28(1), 40–68. doi:10.1214/12-STS395

Rao, J. N. K. (2003). Small Area Estimation. New Jersey: Wiley.

Royall, R. M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. Biometrika, 57(2), 377–387.

Shannon, C. E. (1948). A mathematical theory of communications. Bell System Technical Journal, 27, 379–423.

Tarozzi, A., & Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. The Review of Economics and Statistics, 91(4), 773–792. doi:10.1162/rest.91.4.773

Tobler, W. R. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. Journal of the American Statistical Association, 74(367), 519–530.

[i] See https://ec.europa.eu/eurostat/web/income-and-living-conditions/overview for details.
[ii] In the cases of Portugal, their databases can be found in the international database IPUMS from the University of Minnesota.
[iii] According to EU-SILC, this variable classifies the main job as self-employed with employees, self-employed with no employees, employee and family worker.
[iv] See https://www.ine.es/jaxiT3/Tabla.htm?t=10849&L=1 for details.