REGIONAL SOCIOECONOMIC ASSESSMENTS WITH A GENETIC ALGORITHM. AN APPLICATION ON INCOME INEQUALITY ACROSS MUNICIPALITIES

Elisa Aracil[1,2] , *Elena Maria Diaz*[1], *Gonzalo Gómez-Bengoechea*[1], *Rosalía Mota*[3] & *David Roch-Dupré*[1,2]

[1] Faculty of Economics and Business Administration (ICADE), Universidad Pontifcia Comillas, Madrid, Spain

[2] Institute for Research in Technology (IIT), Universidad Pontifcia Comillas, Madrid, Spain

[3] Faculty of Humanities and Social Sciences, Universidad Pontifcia Comillas, Madrid, Spain

Corresponding author: Gonzalo Gómez-Bengoechea (gonzalo.gomez@comillas.edu)

Research Article

Keywords: Income inequality; Inequality of opportunities; Genetic algorithm; Socioeconomic indicator; Data scarcity; Municipalities

Version of Record: A version of this preprint was published at Social Indicators Research on April 14th, 2024. See the published version at https://doi.org/10.1007/s11205-024-03345-4

Code available at: https://github.com/droch-upco/BUTD-Methodology.

REGIONAL SOCIOECONOMIC ASSESSMENTS WITH A GENETIC ALGORITHM. AN APPLICATION ON INCOME INEQUALITY ACROSS MUNICIPALITIES

Abstract

Available data to depict socioeconomic realities are often scarce at the municipal level. Unlike recurring or continuous data, which are collected regularly or repeatedly, nonrecurrent data may be sporadic or irregular, due to significant costs for their compilation and limited resources at municipalities. To address regional data scarcity, we develop a bottom-up top-down methodology for constructing synthetic socioeconomic indicators combining a genetic algorithm and regression techniques. We apply our methodology for assessing income inequalities at 178 municipalities in Spain. The genetic algorithm draws the available data on circumstances or inequalities of opportunities that give birth to income disparities. Our methodology allows to mitigate the shortcomings arising from unavailable data. Thus, it is a suitable method to assess relevant socioeconomic conditions at a regional level that are currently obscured due to data unavailability. This is crucial to provide policymakers with an enhanced socioeconomic overview at regional administrative units, relevant to allocating public service funds.

Keywords: Income inequality; Inequality of opportunities; Genetic algorithm; Socioeconomic indicator; Data scarcity; Municipalities

1. Introduction

Technological developments and the evolution of political, cultural, and environmental contexts are giving rise to continuously changing socioeconomic realities that demand close monitoring by policymakers and economic agents. This has elicited the creation of numerous social indicators (Espina & Arechávala, 2013; González et al., 2011; Somarriba & Pena, 2009) for different regions to enable optimal decision-making to maximise social welfare. However, the scarcity of socioeconomic data, which is particularly pressing for local administrative units, even in developed countries (Bannor & Oppong-Kyeremeh, 2018; Dan & Lanjouw, 2021; Pandeya et al., 2016; Sanogo, 2019), obscures the identification of social grand challenges in 'left behind places' (Pike et al., 2023).

The issue of socioeconomic data scarcity manifests in various forms. Not only may data be lacking, but existing data may reflect past conditions (as with lagging indicators), encounter publication delays, or be unavailable at a high frequency (Dang et al., 2019). This is even more consequential for composite indicators on socioeconomic realities that rely on the timely availability of base indicators for their estimation. At the municipal scale, critical socioeconomic data are frequently either non-existent or published at irregular frequencies due to the significant administrative costs of their compilation (Bannor & Oppong-Kyeremeh, 2018; Sanogo, 2019). Consequently, there is a lack of recurrent socioeconomic indicators at this granularity, contributing to the scarcity of research below the national or regional level (Aiyar & Ebeke, 2020; Brunori et al., 2018; Checchi et al., 2016; Marrero & Rodríguez, 2012; Ramos & Van de gaer, 2021) and the barriers faced by policymakers in allocation decisions towards municipalities (Bennett & Lemoine, 2014; Millar et al., 2018).

Although extant literature offers various correction methods for addressing missing data or specific data points absent within a dataset (Brunori et al., 2018, 2022), it remains challenging to effectively address the issue of nonrecurrent data, which is non consistently or regularly collected. This paper presents a bottom-up top-down (BUTD) methodology to assess social challenges in data-scarce contexts. This approach enables the development of synthetic socioeconomic indicators using a genetic algorithm and

regression techniques when the required data are unavailable. We apply the BUTD methodology to develop a recurrent indicator of relative income inequality, generated recursively with real-time data. While income inequality is commonly assessed through well-established indicators such as the GINI or the 80/20 poverty ratio, such measurements are seldom recurrently available at the municipal level. Measuring income inequality across local administrative units is highly relevant, given that municipalities rely primarily on central and regional government transfers (Boulant et al., 2016) to deliver public social services, specifically aimed at tackling inequality.

Given that there is a wide set of exogenous circumstances that matter for opportunities in life (Dang, 2014; De Barros et al., 2009; Hick, 2016; Robeyns, 2017; Sen, 1999; World Bank, 2005), we use our methodology to estimate a synthetic indicator by aggregating recurrent data on the unequal opportunities or circumstances that give birth to income disparities. Consequently, our synthetic indicator sheds light on existing income inequality across municipalities and its underlying circumstances. Social public services seek to equalise these 'inequality of opportunities' to narrow income inequality *ex ante* by acting on the drivers of inequality rather than on the outcomes (Fleurbaey & Peragine, 2013; Kovacic et al., 2021; Roemer & Trannoy, 2016). Therefore, our methodology proposes an equal opportunity policy as a set of allocation rules that maximise advantages for the worst-off households, as Ferreira and Gignoux (2011) suggested.

In our application, we recursively estimate a synthetic measure of relative income inequality for 178 municipalities in the Madrid region of Spain and obtain a strong correlation between the benchmark indicator and the synthetic indicator, highlighting the robustness of our methodology. We also find that gender inequality, low work insertion for foreigners, a higher population dispersion, and scarce educational resources are crucial circumstances in explaining income inequality across the municipalities of Madrid.

We contribute to the literature on social indicators (Espina & Arechávala, 2013; González et al., 2011; Somarriba & Pena, 2009) in several ways. First, methodologically, we

propose the development of synthetic indicators by implementing a methodology that combines the genetic algorithm with panel data regression techniques. This approach addresses the challenges of data scarcity and nonrecurrence more effectively than traditional methods of imputation (Dong and Peng, 2013) or forecasting techniques (Taylor and Letham, 2018). Also, we extend the applicability of genetic algorithms and their iterative optimisation process (Holland, 1975) to data-scarce contexts. Second, we also contribute to the local and regional studies literature (Kyriacou et al., 2017; Martínez-Galarraga et al., 2015; Pike et al., 2023; Silveira & Azzoni, 2011) by addressing the issue of limited socioeconomic data at municipal administrative units. Moreover, we extend the empirical literature on income inequality (Bourguignon, 2017) and inequality of opportunities (Aaberge et al., 2011; Bourguignon et al., 2007; Brunori et al., 2013; Ferreira & Gignoux, 2011; Fleurbaey & Peragine, 2013; Lefranc et al., 2008) by providing a method that allows to consider the multidimensionality of circumstances that underlie income inequality.

Finally, from the policymaking perspective, we provide a tool for social public services to rectify these 'inequalities of opportunities' in order to proactively reduce income inequality by addressing the root causes rather than the outcomes (Fleurbaey & Peragine, 2013; Kovacic et al., 2021; Roemer & Trannoy, 2016). Thus, we propose an equal opportunity policy tool comprising allocation rules designed to maximize benefits for the most disadvantaged households, aligning with the recommendations of Ferreira and Gignoux (2011).

The paper proceeds as follows. Section 2 describes the BUTD methodology for assessing socioeconomic conditions. Section 3 illustrates the methodology by depicting relative income inequality and its underlying circumstances in 178 municipalities of Madrid. Finally, Section 4 presents the conclusions and implications of our findings for scholars and policymakers.

2. The bottom-up top-down methodology

The BUTD methodology addresses the issue of data scarcity when key socioeconomic indicators are not available at the desired frequency, that is, when they are nonrecurrent. Nonrecurrence occurs when there is no systematization in the compilation of indicators, common for experimental datasets, datasets deriving from low-fund projects, or whenever unexpected conditions prevent data collection and subsequent reporting.

The complexities associated with nonrecurrent data differ from those associated with missing data. Under nonrecurrent data, the difficulty in identifying the missing data patterns – classified as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Rubin, 1976)– and the large amount of missing data render traditional imputation techniques less effective. Furthermore, forecasting methods can be challenging to apply in contexts of nonrecurrent data due to the insufficiency of data to generate accurate predictions. In response to these concerns, we propose the 'BUTD' methodology, a hybrid approach that merges an optimization algorithm with panel data regression techniques. This methodology effectively tackles the pronounced issues of data scarcity and nonrecurrence, which are limitations that conventional methods fail to address satisfactorily. In the first phase (bottom-up), we aggregate nonrecurrent data (X_{NR}) to develop a benchmark indicator (Y_{NR}), which is, therefore, nonrecurrent as well. Traditional methods for developing composite indicators are implemented at this stage. In the second phase (top-down), we implement a genetic algorithm together with regression techniques to construct a recurrent synthetic indicator (Y_R) that mimics the benchmark indicator by using recurrently available socioeconomic data (X_R). The procedures are detailed in the next sections.

2.1 Bottom-Up

The bottom-up phase consists of the standard aggregation process for estimating composite indicators (see Gan et al., 2017, for a review). This includes the selection of the base indicators X_{NR} (nonrecurrent by nature), their normalisation, the assignation of weights, ω_{NR}, for each indicator, and the determination of the aggregation method.

In our application for income inequality, we select the base indicators and their respective weights following the Budget Allocation Process (BAP), based on expert criteria. This method is suitable in the presence of a limited number of base indicators -no more than 10- (OECD, 2008). Then we apply a normalization of the base indicators (dividing the whole data series of each indicator by its maximum) in order to homogenize the different units of measurement associated with the indicators to a common scale in the range (0-1]. Finally, for the aggregation method, we assume limited substitutability or trade-off between the base indicators, which requires the application of a geometric aggregation (Lafortune et al., 2018).

The resulting benchmark socioeconomic indicator Y_{NR} is then defined as:

$$Y_{NR} = \prod({X_{NR}}^{\omega_{NR}}) \quad (1)$$

where the values of Y_{NR} range from 0 to 1 and higher values represent worse-off states of income inequality between regions.

2.2 Top-Down

The top-down phase in the BUTD methodology consists of building a synthetic indicator (Y_R) using recurrently available indicators (X_R). In this case, determining the weights (ω_R) using expert criteria is rendered problematic due to the large number of recurrent indicators. Therefore, we cannot apply the standard aggregation process as in the bottom-up phase.

Because recurrent indicators denote circumstances that relate to the intended socioeconomic condition, we can estimate the relationship between these circumstances and the benchmark *(*Y_{NR}) as follows:

$$Y_{NR} = \alpha + \omega_R X_R + \varepsilon$$

$$s.t. \tag{2}$$

$$\omega_R \geq 0$$

where α is a constant vector and ε is a vector of error terms.

The weights ω_R are restricted to be nonnegative since recurrent indicators relate positively to the condition represented by the benchmark indicator *(*Y_{NR}). Therefore, while ordinary least squares (OLS) could be implemented to estimate the weights ω_R, including all available indicators in an OLS estimation would result in failure to comply with this restriction due to the existing collinearity between the indicators X_R.

Alternatively, we can select a subset of recurrent indicators, X_R^*, such that the correlation between the selected indicators is low enough to obtain solely positive weights when regressing income inequality Y_{NR} against X_R^*. This complies with the nonnegative restriction and ensures a parsimonious model that avoids redundant information and overfitting; however, it poses a highly nonlinear optimisation problem that requires optimisation algorithms.

2.2.1 A genetic algorithm to select recurrent indicators

We use a genetic algorithm (Holland, 1975) to select the indicators from X_R that should be included in the subset X_R^* to provide the best-fit model for Y_{NR}. The genetic algorithm is a nature-inspired optimisation algorithm that runs an iterated selection to ensure the best 'fit' or correspondence to the objective or 'fitness' function. These algorithms have demonstrated great flexibility and success in dealing with computationally intensive, highly nonlinear, and nonconvex problems (Ertenlice & Kalayci, 2018; Mavrovouniotis et al., 2017). The genetic algorithm, first introduced to economics by Miller (1996), has had a wide range of applications, such as optimisation in operations management or carbon

emissions (Dat et al., 2012; Kia et al., 2014; Lee, 2018; Mazinani et al., 2013; Soleimani et al., 2017), scheduling (Efthymiou et al., 2017; Kuo et al., 2016), business planning (Shin & Lee, 2002) and financial trading systems (Dempster & Jones, 2001). Moreover, Diaz and Perez-Quiros (2021) implemented the genetic algorithm in selecting economic indicators. However, to our knowledge, no previous attempts to use the genetic algorithm to develop synthetic socioeconomic indicators exist. Nonetheless, this algorithm is ideal for our application, given that its combinatorial (Morini & Pellegrino, 2018) and binary nature is well suited for the selection of indicators.[1]

We apply the genetic algorithm to a universe of 2^{40} different combinations of socioeconomic indicators. The algorithm evaluates, selects, cross breeds, and mutates the best alternatives over several iterations until a stopping criterion is met, and it converges to the optimal combination of recurrent indicators X_R^* as a solution to the model in Equation 2. The technical details of how this is performed can be found in Appendix A.

2.2.2 Definition of the Synthetic Indicator

We are now able to construct a synthetic recurrent indicator Y_R, which is defined as an arithmetic aggregation[2] of the selected indicators, X_R^*, as follows:

$$Y_R \equiv \alpha^* + \omega_R^* X_R^*, \tag{3}$$

where the parameters α^* and ω_R^* are estimated through OLS. Similar to the benchmark indicator Y_{NR}, the greater the value of Y_R is, the greater the relative income inequality.

---

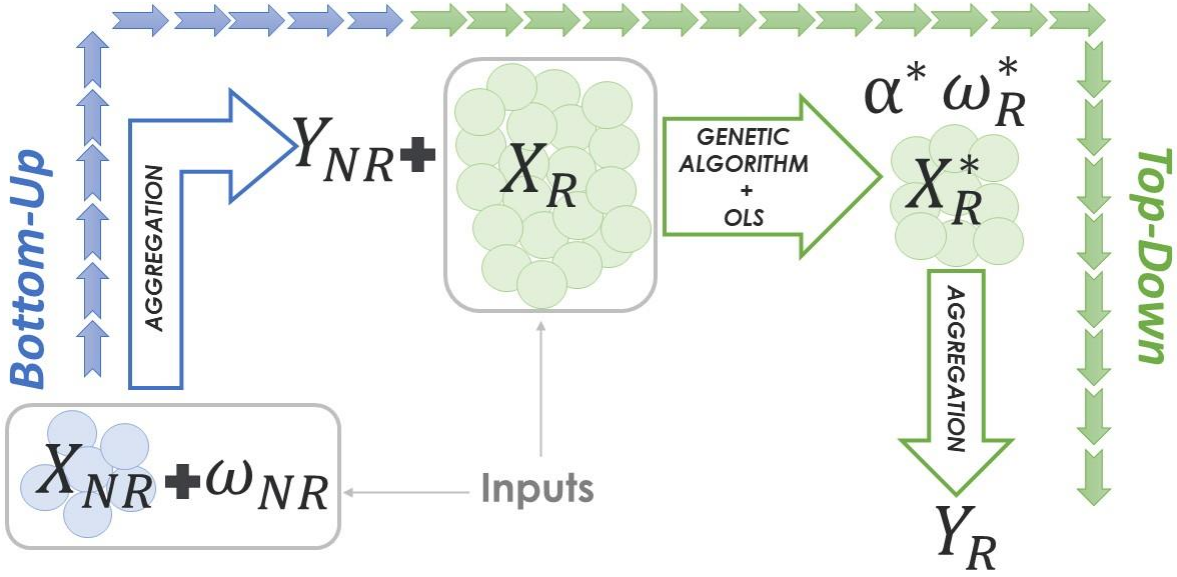
[1] Other nature-inspired optimization algorithms could have been used with similar results (especially those ones with binary structures). Some examples where the genetic algorithm has been compared with other nature-inspired optimization algorithms (such as the Particle Swarm Optimization Algorithm and the Fireworks Algorithm) in the context of complex problems of optimal selection are Roch et al., (2021 a; 2021 b). These studies offer very similar results with the different algorithms, with a slight outperformance of the genetic algorithm.

[2] Although we have selected an arithmetic aggregation to estimate Y_R, the BUTD methodology can be extended to cases in which other aggregation methods may be preferred. We have opted for the arithmetic aggregation due to the low level of substitutability (Lafortune et al., 2018) across the indicators within X_R^* and the similarity between formulas of the OLS regression (Equation 2) and the arithmetic aggregation. The selected aggregation method determines the model in Equation 2.

By construction, these estimates result in the highest R^2-statistic, subject to restrictions. However, our methodology does not intend to determine a causal relationship between the circumstances underlying inequality, X_R^*, and income inequality, Y_{NR}, which may be endogenous (Ramos & Van de gaer, 2021). In contrast, we develop a synthetic indicator on income inequality according to its underlying disparities in opportunities, where the estimated weights ω_R^* signal the relevance of each circumstance with respect to overall income inequality. The error term ε captures other inequalities in opportunities that may arise in different contexts and the differences in effort as a nonmeasurable determinant of income inequality (Ramos & Van de gaer, 2016). Figure 1 summarises the bottom-up top-down methodology in a flowchart. Replication materials are located in a public repository[3].

Figure 1

Flowchart of the Bottom-Up Top-Down Methodology



[3] The link to the repository will be open access from the publication date onwards.

2.2.3 Recursive Estimation of the Synthetic Indicator

We apply a recursive or real-time estimation process to ensure the appropriateness of our methodology, ensuring that only available information at the time of estimation is utilized both to assign the weights as well as perform the selection of the recurrent indicators.[4] Note, therefore, that the genetic algorithm uses a set of information available at time t, denoted I_t, which contains all historical values of the benchmark indicator, $Y_{NR,t-1,\dots t-T}$, and all historical and current values of the recurrent indicators, $X_{R,t,\dots t-T}$, where T denotes the number of periods in the sample. In this sense, the selection of recurrent indicators and determination of weights will depend only on the historical values of $Y_{NR,t-1,\dots t-T}$ and $X_{R,t-1,\dots t-T}$, while current values of $X_{R,t}$ will be used to construct the synthetic indicator $Y_{R,t}$ at time t.

3. An empirical application for income inequality in Madrid municipalities

We implement the BUTD methodology to construct a synthetic indicator of relative income inequality for 178 municipalities[5] in the region of Madrid, Spain.

Madrid forms the largest regional economy in Spain and within the 15 highest European regions in GDP per capita in 2020, according to Eurostat data. Notwithstanding, it holds high levels of income inequality across municipalities, rendering it of particular interest for our study. In addition, Madrid is one of the most populated regions in Europe, for which it is considered a predominantly urban region (Eurostat, 2018). This guarantees sufficient homogeneity across municipalities for the validity of our model, as opposed to regions where population dispersion leads to significant variations in the relationship between disparities of opportunities and income inequality. Finally, its administrative

[4] A real-time estimation restricts input data to information available at the time of estimation. In our application, this implies that selection and weights are assigned using only 2015 data, and those are applied to 2016 recurrent indicators to build that year's synthetic indicator for inequality. Likewise, only data for 2015 and 2016 are used to assign new selection and weights that are then applied to 2017 recurrent indicators for the estimation of that year's synthetic indicator.

[5] The municipality of Madrid city is not included in our analysis due to its distinctive features, which would require a particular and adapted analytical approach for municipalities with larger populations (Brezzi et al., 2011; Royuela et al., 2014). Madrid represented 50% of the region's population and 55% of the region's total economic activity in 2020, according to INE, which makes it an outlier in our analysis.

organisation into 178 municipalities provides a large and representative sample for testing our methodology.

Given the availability of the data, we restrict our analysis to the period between 2015 and 2017.

3.1. Bottom-Up. Development of a benchmark income inequality indicator

We begin by following the bottom-up phase of the BUTD methodology described in Section 2.1, which consists of selecting the base indicators and their corresponding weights (Equation 1), according to expert criteria.

We first define X_{NR} to contain the Gini coefficient and the 80/20 poverty ratio as key base indicators for income inequality. The Gini coefficient is frequently used as a measure of inequality across the entire income distribution of a population (Kakwani, 1980; Nygård & Sandström, 1989; Yitzhaki & Schechtman, 2013), whereas the 80/20 poverty ratio is the quotient between the income of the lowest and highest quintiles of the population as an estimate of the spread within this distribution (Banerjee et al., 2021; Lustig et al., 2013; Lustig, 2018). Therefore, their combination is value-enhancing since the Gini coefficient captures changes in the overall income distribution, while the 80/20 poverty ratio better represents shifts occurring in the highest and lowest income levels.

Nonetheless, both the Gini coefficient and the 80/20 poverty ratio show income inequality *within* a municipality, whereas our goal is to identify relative income inequality *across* municipalities. To enable income inequality assessments inter-municipalities, we include GDP per capita in X_{NR}, which is inverted to ensure unambiguity [6].

[6]Unambiguity ensures a homogeneous interpretation of the indicators' performance (increments or decrements). In our application, the higher the value of an indicator, the more vulnerable the municipality.

The weights $\omega_{NR} = [0.35,\ 0.35,\ 0.30]$ are then defined for the Gini coefficient, the 80/20 poverty ratio, and the inverse of GDP per capita. Based on the BAP methodology, expert criteria have selected these weights to keep the dominance of the core inequality indicators, Gini and the 80/20 poverty ratio. GDP per capita can be regularly obtained from the Almudena database from Spain's Instituto Nacional de Estadística (INE), however, the Gini coefficient and the 80/20 poverty ratio for municipalities frequently suffer from limited availability and discontinuity problems (Wilkinson & Pickett, 2009). For the municipalities of Madrid, these are only available for 2015 to 2017 in the Atlas experimental dataset of Spain's INE. Therefore, we can only construct for this period.

All three base indicators are normalised and geometrically aggregated, as described in Section 2.1, to obtain Y_{NR}. Note that given the range between 0 and 1 for the values of Y_{NR}, we can then interpret that a hypothetical municipality would reach a value of 1 if it had the lowest GDP per capita in the region of study and the highest income inequality according to the Gini coefficient and the 80/20 poverty ratio.

Table 1 presents the estimates of the mean benchmark income inequality Y_{NR} for each decile of the municipalities of Madrid from 2015 to 2017 and the mean values for its nonrecurrent base indicators, that is, the Gini coefficient, the 80/20 poverty ratio, and GDP per capita. Estimations show significant income inequality across the municipalities in Madrid, with values for the benchmark indicator ranging from close to 0.4 in the lowest decile to more than 0.7 in the highest decile. By construction, the benchmark indicator is directly proportional to the Gini coefficient and the 80/20 poverty ratio and inversely proportional to GDP per capita, with no significant outliers in the sample. Due to its construction methodology, the benchmark indicator should not be interpreted as an intertemporal inequality variable since we normalise each component with respect to annual maximums. Finally, our study period has no temporal effects, with consistent means for all indicators across the sampled years.

If the GDP is not inverted, its interpretation would be the opposite (i.e., the higher the value, the less vulnerable the municipality).

Table 1
Summary statistics.

Decile	Year	Benchmark indicator	Nonrecurrent base indicators		
			Gini coefficient	80/20 ratio	GDP per capita
10	2015	0.720	37.453	3.565	11,496.353
	2016	0.704	37.082	3.524	11,592.471
	2017	0.752	35.300	3.306	11,663.000
9	2015	0.658	35.359	3.265	13,202.471
	2016	0.651	35.453	3.318	13,232.941
	2017	0.688	34.106	3.076	13,799.529
8	2015	0.632	34.788	3.088	13,909.353
	2016	0.617	34.512	3.059	13,945.059
	2017	0.657	32.859	2.882	14,298.647
7	2015	0.612	34.876	3.129	15,831.000
	2016	0.609	35.229	3.212	15,934.647
	2017	0.643	33.629	2.947	16,291.059
6	2015	0.586	34.329	3.053	17,365.529
	2016	0.568	34.341	3.012	18,139.353
	2017	0.609	33.059	2.871	18,626.588
5	2015	0.562	33.018	2.900	17,991.059
	2016	0.553	33.294	2.941	18,689.176
	2017	0.594	32.265	2.788	19,279.294
4	2015	0.546	32.735	2.829	19,079.176
	2016	0.538	32.829	2.876	19,394.824
	2017	0.573	31.500	2.724	20,189.647
3	2015	0.518	32.718	2.959	24,026.706
	2016	0.507	32.582	2.959	24,569.706
	2017	0.531	31.276	2.741	26,308.824
2	2015	0.484	32.100	2.775	27,150.125
	2016	0.481	32.331	2.813	27,408.188
	2017	0.508	31.194	2.631	28,934.063
1	2015	0.421	32.613	2.888	48,327.750
	2016	0.414	32.950	2.913	49,420.313
	2017	0.445	31.531	2.738	48,844.188

Note: Mean values by year for each decile group of municipalities. The grouping by deciles has been made according to the ranked values of the benchmark indicator (Y_{NR}).

3.2. Top-Down. Development of a synthetic income inequality indicator

We proceed to construct the synthetic income inequality indicator for the municipalities of Madrid over the 2015–2017 period, following the top-down phase of the BUTD methodology. We selected recurrent and standard indicators (X_R) that are available annually in the Almudena database of Spain's Instituto Nacional de Estadística (INE), representing unequal circumstances that determine inequality across multiple dimensions (Alberti et al., 2021; World Bank, 2005). We classify X_R according to extant literature (i.e., Cabrera et al., 2021; Espina & Arechávala, 2013; González et al., 2011; Somarriba & Pena, 2009) into the categories of demography, labour market structures, income, and living conditions[7]. The selected indicators, commonly found in the literature, originate from the same source (INE), correspond to municipalities within the same region (Madrid), and are built under the same standards, which guarantees their validity and comparability, even if they are not available with the same regularity. Appendix B describes the 39 indicators in X_R, including its categorisation, definition, and primary source, while Appendix C contains the descriptive statistics of both recurrent and nonrecurrent indicators.

Each recurrent indicator (X_R) is divided by its maximum (for re-scaling) and, when needed, inverted to maintain unambiguity. Then, they are used as input for optimising the model described in Equation 2. Within X_R, a subset X_R^* will be optimally selected by the genetic algorithm to represent overall income inequality, using only the information available at each point in time (I_t).

[7] These categories are merely indicative and different groupings do not affect the estimations.

3.3. Results for the synthetic indicator of income inequality across municipalities

The selected recurrent indicators X_R^*, their corresponding weights ω_R^*, and the intercept α^* determined by the BUTD methodology for the municipalities of Madrid are presented in Table 2. For parsimony, we present the estimated results using all information available until 2017, I_{2017}.[8]

Table 2

Recurrent indicators and weights selected by the genetic algorithm $(\alpha^*, \omega_R^*, X_R^* | I_{2017})$

Recurrent indicators on circumstances (X_R^*)	Demography	Labour market	Income	Living conditions
Female population	**0.3528****			
Senior population	0.0041			
Dependency ratio	0.0617			
Foreign population	0.0435			
Foreign female population	0.0330			
Foreign working population		**0.0561***		
Young working population		0.0197		
Foreigners' unemployment		0.0420		
Female work insertion		**0.0751****		
Foreign extra-EU work insertion		0.0051		
GDP per capita			**0.4401*****	
Number of tax declarations			0.0489	
Families with minimum insertion income			0.0097	
Electricity consumption				0.0010
Water consumption				0.0377
Population dispersion				**0.5873*****
Students per school unit				**0.0574****
Intercept	-0.1460			
No. of observations	534			
R^2	0.746			

***p < 0.01; **p < 0.05; *p < 0.1

[8] Results estimated with I_{2015} and I_{2016} are available upon request.

The genetic algorithm selects 17 recurrent indicators for the model described in Equation 2 from all four categories, effectively representing all relevant dimensions for income inequality. The estimated model explains up to 74.56% of the variance of income inequality across the municipalities of Madrid, whereas the unexplained variance corresponds to differences in unobserved circumstances and efforts.

We can also observe statistically significant correlations between income inequality and underlying circumstances for specific indicators across each category. These include the foreign female population, which signals the relevance of gender inequality as a precursor of income inequality, in line with prior studies (Aaberge & Brandolini, 2015; Cabrera et al., 2021; Ramos & Van de gaer, 2016). This is also reflected in the significance of female work insertion, which implies that municipalities in which tightening labour markets hinders the generation of new work contracts for women have higher levels of income inequality, as highlighted at a supranational level by Marrero and Rodríguez (2012).

Additionally, macroeconomic conditions are significantly relevant for determining income inequality in the municipalities of Madrid, as signalled by the weight assigned to GDP per capita. This effectively shows that prevailing income levels are part of the opportunity sets underlying income inequality (Hufe et al., 2018; Lefranc et al., 2008, 2009). Therefore, as in the case of the estimation of Y_{NR}, the genetic algorithm selects GDP per capita as an indicator of differences in mean income level across territories.

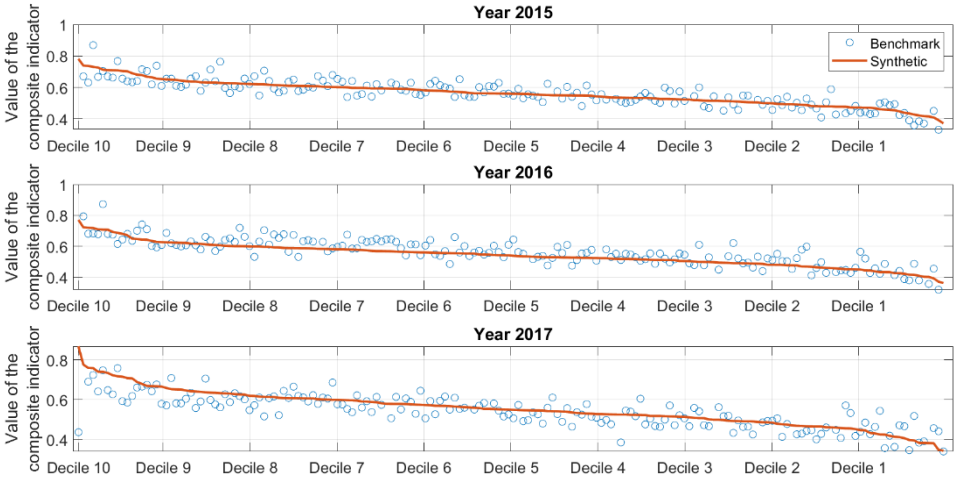
In addition, we find a significant relationship between income inequality and population dispersion, suggesting that municipalities with greater population dispersion, classified as rural, suffer from higher levels of income inequality (Brock, 2020; Niehues & Peichl, 2014). Finally, there is also a strong link between the resources devoted to education and income inequality in a municipality, which highlights the influence of the availability of public services on income inequality (Cabrera et al., 2021; Chatterjee & Turnovsky, 2012; World Bank, 2005).

We caution that a weight that is not significantly different from zero for a selected indicator does not imply a null relevance in relation to income inequality. These base indicators are still selected by the genetic algorithm, for which their inclusion necessarily improves the explanation of income inequality across the municipalities of Madrid.

Followingly, based on the genetic algorithm's selected indicators X_R^* and optimal parameters ω_R^* and α^* given the information set I_t for each time period, we can recursively estimate a recurrent synthetic indicator of income inequality Y_R for the municipalities of Madrid with Equation 3. The results are shown in Figure 2, which ranks the municipalities in Madrid in descending order of relative income inequality according to the synthetic indicator from 2015 to 2017. Synthetic values are plotted along with the benchmark values for comparison. Figure 2 shows that the synthetic indicator closely tracks the benchmark values of income inequality, thereby providing evidence for the BUTD methodology as a suitable vehicle to assess relative income inequality in the presence of data scarcity. The correlation between the benchmark and synthetic indicators are 85.51%, 84.30% and 76.92% for 2015, 2016 and 2017, respectively. Again, no significant temporal effects can be observed. Given that for year 2015 there is no historical data available, estimated coefficients and selected indicators represent an in-sample rather than a real-time estimation.

Figure 2.

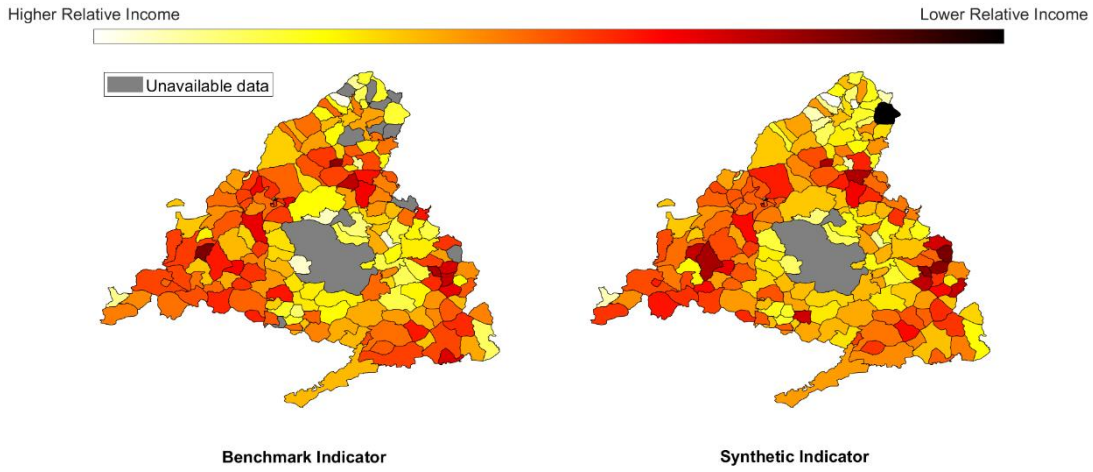
Benchmark and synthetic indicators for income inequality across Madrid municipalities



Results can also be visualised geographically in Figure 3, where municipalities in darker colours present a lower relative income with respect to municipalities in lighter colours for 2017.[9] These disparities signal the optimal social services funding allocation required for narrowing income inequality in the region.

Figure 3.

Mapping benchmark and synthetic indicators for income inequality across Madrid municipalities (year 2017)



Figures 2 and 3 also depict the robustness of our methodology for the development of a synthetic indicator that accurately replicates relative income inequality across municipalities.

[9] A mapping visualization of relative income inequality during 2015 and 2016 depicts similar results. These are available upon request.

4. Conclusions

At a municipal level, key socioeconomic data are either unavailable or irregularly released, hindering a precise picture of local socioeconomic conditions. This paper presents the bottom-up-top-down (BUTD) methodology, which generates a synthetic indicator on a given socioeconomic reality whenever facing data scarcity. The methodology integrates a genetic algorithm with regression techniques to select and weight recurrently available indicators, thereby generating a synthetic indicator that closely replicates a nonrecurrent benchmark measure.

To illustrate our methodology, we focus on income inequality across municipalities, where data scarcity hampers the adequate allocation of public service funding and policymaking (Aiyar & Ebeke, 2020; Checchi et al., 2016; Marrero & Rodríguez, 2012; Ramos & Van de gaer, 2021). Specifically, we assess relative income inequality across 178 municipalities from Madrid, Spain's most significant economic region.

The genetic algorithm achieves a strong correlation between the benchmark and synthetic indicators, which supports the robustness of our methodology for estimating income inequality in the presence of nonrecurrent data. The resulting synthetic indicator identifies a set of circumstances that underlie disparities in income, specifically gender inequality, female and foreign unemployment, rurality or population dispersion, and public education resources. These results provide empirical support to prior literature (Bourguignon et al., 2007; Ferreira & Gignoux, 2011; Ferreira & Peragine, 2016; Lefranc et al., 2008; Roemer, 1993, 2000; Roemer & Trannoy, 2016).

This study is not without limitations. First, our application is constrained by the availability of a limited number of nonrecurrent indicators to construct the benchmark that the BUTD synthetically replicates. This issue may overlook other crucial factors for measuring inequalities as a benchmark. However, this limitation pertains to our specific application but does not extend to our methodology, which enables the replication of existing nonrecurrent data regardless its accuracy. Second, our findings are limited to the socioeconomic conditions across our sample, and, thus not generalisable to other regions. Nonetheless, the results serve as an exemplary overview of how the

methodology can be applied and solve the problem of data scarcity in sub-national administrative units.

Our approach offers policymakers an instrument to assess social issues that lack recurrent data, thereby facilitating the allocation of public funds for their mitigation. This is especially valuable for decentralised government structures where national or subnational administrative units perform revenue collection, while basic social services provision decisions are transferred to municipalities. In our application of income inequality among municipalities, our findings suggest that policymakers should concentrate on certain key conditions such as gender inequality, integration of foreign workers, population distribution, and educational resources. Providing methodologies for assessing life conditions may promote quality of life (Shek & Wu, 2018) and catalyze an optimal, transparent and fair distribution of public funds.

Because the proposed methodology is designed to address the issue of data scarcity, the improvement of local data systems would make our contribution redundant. However, although enhanced data availability is a desirable path that public administrations will eventually walk through in the upcoming years, it will not be the case in many regions in the near future. Therefore, our methodology will continue bridging this gap for policymakers and society. Future researchers could implement our methodology using alternative optimisation algorithms such as binary particle swarm optimisation (Mirjalili & Lewis, 2013). Our methodology can be extended from municipalities to other units of analysis and from the problem of nonrecurrent data to other data issues (i.e., lagging indicators or differences in frequencies). Finally, our method opens future research avenues by extending its application to assess other socioeconomic conditions, such as poverty or specific forms of poverty like energy poverty, in presence of nonrecurrent data.

References

Aaberge, R., & Brandolini, A. (2015). Multidimensional poverty and inequality. In *Handbook of income distribution*, 2, 141-216. Elsevier.

Aaberge, R., Mogstad, M., & Peragine, V. (2011). Measuring long-term inequality of opportunity. *Journal of Public Economics*, 95(3), 193-204.

Aiyar, S. & Ebeke C. (2020). Inequality of opportunity, inequality of income and economic growth. *World Development*, 136, 105115.

Alberti, V., Banys, K., Caperna, G., Del Sorbo, M., Fregoni, M., Havari, E., Kovacic, M., Lapatinas, A., Litina, A., Montalto, V. , Tacao Moura, CJ, Neher, F., Panella, F., Peragine, V., Pisoni, E., Stuhler, J., Symeonidis, K., Verzillo, S. y Boldrini, M., (2021). Monitoring Multidimensional Inequalities in the European Union. *Joint Research Centre*, EUR 30649 EN, JRC123911. Publications Office of the European Union, Luxembourg.

Banerjee, A., Duflo, E., & Sharma, G. (2021). Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, 3(4), 471-486.

Bannor, R. K., & Oppong-Kyeremeh, H. (2018). Extent of poverty and inequality among households in the Techiman municipality of Brong Ahafo region, Ghana. *Journal of Energy and Natural Resource Management,* 1(1), 26-36.

Bennett, N., & Lemoine, G. J. (2014). What a difference a word makes: Understanding threats to performance in a VUCA world. *Business Horizons*, 57(3), 311-317.

Boulant, J., Brezzi, M., & Veneri, P. (2016). Income levels and inequality in metropolitan areas: A comparative approach in OECD countries. *OECD Regional Development Working Papers*, No. 2016/06, OECD Publishing, Paris.

Bourguignon, F., Ferreira, F.H.G. & Walton, M. (2007). Equity, efficiency and inequality traps: A research agenda. *The Journal of Economic Inequality*, 5, 235–256.

Bourguignon, F. (2017). Global inequality. In: *The Globalization of Inequality*, 9-40. Princeton University Press.

Bouzarovski, S., & Tirado-Herrero, S. (2017). The energy divide: Integrating energy transitions, regional inequalities and poverty trends in the European Union. *European Urban and Regional Studies*, 24(1), 69-86.

Brezzi, M., L. Dijkstra & V. Ruiz (2011). OECD extended regional typology: The economic performance of remote rural regions. *OECD Regional Development Working Papers*, No. 2011/06.

Brock, J. M. (2020). Unfair inequality, governance and individual beliefs. *Journal of Comparative Economics*, 48(3), 658-687.

Brunori, P., Hufe, P., & Mahler, D. G. (2018). The roots of inequality: Estimating inequality of opportunity from regression trees. *World Bank Policy Research Working Paper*, No. 8349.

Brunori, P., Ferreira, F. H., & Peragine, V. (2013). Inequality of opportunity, income inequality, and economic mobility: Some international comparisons. In Getting development right: Structural transformation, inclusion, and sustainability in the post-crisis era (pp. 85-115). New York: Palgrave Macmillan US.

Brunori, P., Salas-Rojo, P., & Verne, P. (2022). Estimating inequality with missing incomes. London School of Economics. International Inequalities Institute. *Working Paper*, No. 82.

Cabrera, L., Marrero, G.A., Rodríguez, J.G., & Salas-Rojo, P. (2021). Inequality of opportunity in Spain: New insights from new data. *Review of Public Economics*, 237(2), 153-185.

Chatterjee, S., & Turnovsky, S. J. (2012). Infrastructure and inequality. *European Economic Review*, 56(8), 1730-1745.

Checchi, D. Peragine, V. & Serlenga, L. (2016). Inequality of opportunity in Europe: Is there a role for institutions? In: *Lorenzo Cappellari & Solomon W. Polachek & Konstantinos Tatsiramos (ed.), Inequality: Causes and Consequences*, 43, 1-44, Emerald Publishing Ltd.

Checchi, D. & Peragine, V. (2010). Inequality of opportunity in Italy. *The Journal of Economic Inequality*, 8, 429-450.

Dang, A.T. (2014). Amartya Sen's capability approach: A framework for well-being evaluation and policy analysis? *Review of Social Economy*, 72(4), 460-484.

Dang, H. A., Jolliffe, D., & Carletto, C. (2019). Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. *Journal of Economic Surveys*, 33(3), 757-797.

Dat, L. Q., Linh, D. T. T., Chou, S. Y., & Vincent, F. Y. (2012). Optimizing reverse logistic costs for recycling end-of-life electrical and electronic products. *Expert Systems with Applications*, 39(7), 6380-6387.

De Barros, R. P., Ferreira, F., Vega, J., & Chanduri, J. (2009). Measuring inequality of opportunities in Latin America and the Caribbean. *World Bank Publications*.

Dempster, M. A. H., & Jones, C. M. (2001). A real-time adaptive trading system using genetic programming. *Quantitative Finance*, 1(4), 397.

Diaz, E. M., & Perez-Quiros, G. (2021). GEA tracker: A daily indicator of global economic activity. *Journal of International Money and Finance*, 115, 102400.

Dong, Y. and Peng, C.J. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 222.

Efthymiou, D., Chrysostomou, K., Morfoulaki, M., & Aifantopoulou, G. (2017). Electric vehicles charging infrastructure location: A genetic algorithm approach. *European Transport Research Review*, 9(2), 27.

Ertenlice, O. & Kalayci, C.B. (2018). A survey of swarm intelligence for portfolio optimization: Algorithms and applications. *Swarm and Evolutionary Computation*, 39, 36–52.

Espina, P. Z., & Somarriba, N. (2013). An assessment of social welfare in Spain: Territorial analysis using a synthetic welfare indicator. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, 111(1), 1-23.

Eurostat (2018). Methodological manual of territorial typologies. Eurostat Statistical Books.

Ferreira, F. H., & Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*, 57(4), 622-657.

Ferreira, F. H., & Peragine, V. (2016). Individual responsibility and equality of opportunity. *The Oxford Handbook of Well-Being and Public Policy*.

Fleurbaey, M., & Peragine, V. (2013). Ex ante versus ex post equality of opportunity. *Economica*, 80(317), 118-130.

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, 31(5), 694-708.

Gan, X., Fernandez, I. C., Guo, J., Wilson, M., Zhao, Y., Zhou, B., & Wu, J. (2017). When to use what: Methods for weighting and aggregating sustainability indicators. *Ecological Indicators*, 81, 491-502.

González, E., Cárcaba, A., & Ventura, J. (2011). The importance of the geographic level of analysis in the assessment of the quality of life: The case of Spain. *Social Indicators Research*, 102, 209-228.

Holland, J.H. (1975) Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. *University of Michigan Press.*

Hick, R. (2016). Material poverty and multiple deprivation in Britain: The distinctiveness of multidimensional assessment. *Journal of Public Policy*, 36 (2), 277-308.

Hufe, P., Kanbur, R. & Peichl, A. (2018) Measuring unfair inequality: Reconciling equality of opportunity and freedom from poverty. *CESifo Working Paper Series* No. 7119.

Jusot, F., Tubeuf, S., & Trannoy, A. (2013). Circumstances and efforts: How important is their correlation for the measurement of inequality of opportunity in health? *Health Economics,* 22(12), 1470-1495.

Kakwani, N. C. (1980). Income inequality and poverty: Methods of estimation and policy implications. *Population and Development Review*, 6, 673.

Kia R., Khaksar-Haghani, F., Javadian, N., Tavakkoli-Moghaddam, R. (2014). Solving a multi-floor layout design model of a dynamic cellular manufacturing system by an efficient genetic algorithm. *Journal of Manufacturing Systems,* 33(1), 218–232.

Kilkiş, Ş. (2016). Sustainable development of energy, water and environment systems index for Southeast European cities. *Journal of Cleaner Production*, 130, 222-234.

Kovacic, M., Verzillo, S. & Peragine, V. (2021). Using survey and administrative data to gain insights on the evolution of inequality of opportunity in the EU. In *Dominguez-Torreiro, M. and Papadimitriou, E. (ed.). Monitoring Multidimensional Inequalities in the European Union*, 75-97. Luxembourg: Publications Office of the European Union.

Kyriacou, A. P., Muinelo-Gallo, L., & Roca-Sagalés, O. (2017). Regional inequalities, fiscal decentralization and government quality. *Regional Studies*, 51(6), 945-957.

Kuo, Y. H., Rado, O., Lupia, B., Leung, J. M., & Graham, C. A. (2016). Improving the efficiency of a hospital emergency department: a simulation study with indirectly imputed service-time distributions. *Flexible Services and Manufacturing Journal*, 28(1-2), 120-147.

Lafortune, G., Fuller, G., Moreno, J., Schmidt-Traub, G., & Kroll, C. (2018). SDG index and dashboards detailed methodological paper. *Sustainable Development Solutions Network.*

Lee C.K.H. (2018). A review of applications of genetic algorithms in operations management. *Engineering Applications of Artificial Intelligence,* 76, 1–12.

Lefranc, A., Pistolesi, N. & Trannoy, A. (2008). Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? *Review of Income and Wealth*, 54 (4), 513–546.

Lefranc, A., Pistolesi, N. & Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics*, 93(11–12), 1189–1207.

Lustig, N., Lopez-Calva, L. F., & Ortiz-Juarez, E. (2013). Declining inequality in Latin America in the 2000s: The cases of Argentina, Brazil, and Mexico. *World Development*, 44, 129-141.

Lustig, N. (2018). Commitment to equity handbook: Estimating the impact of fiscal policy on inequality and poverty. *Brookings Institution Press.*

Marmot, M. (2005). Social determinants of health inequalities. *Lancet*, 365, 1099-1104.

Marrero, G. & Rodríguez, J. G. (2012). Inequality of opportunity in Europe. *Review of Income and Wealth*, 58(4), 597–621.

Martínez-Galarraga, J., Rosés, J. R., & Tirado, D. A. (2015). The long-term patterns of regional income inequality in Spain, 1860–2000. *Regional Studies*, 49(4), 502-517.

Mavrovouniotis, M., Li, C., Yang, S. (2017). A survey of swarm intelligence for dynamic optimization: Algorithms and applications. *Swarm and Evolutionary Computation,* 33, 1–17.

Mazinani, M., Abedzadeh, M., Mohebali, N. (2013). Dynamic facility layout problem based on flexible bay structure and solving by genetic algorithm. *The International Journal of Advanced Manufacturing Technology*, 65(5–8), 929–943.

Millar, C. C. J. M., Groth, O., & Mahon, J. F. (2018). Management innovation in a VUCA world: Challenges and recommendations. *California Management Review*, 61(1), 5–14.

Morini, M., & Pellegrino, S. (2018) Personal income tax reports: A genetic algorithm approach. *European Journal of Operational Research*, 264(3), 994-1004.

Mussida, C., & Parisi, M. L. (2018). Immigrant groups' income inequality within and across Italian regions. *The Journal of Economic Inequality*, 16, 655-671.

Niehues, J., & Peichl, A. (2014). Upper bounds of inequality of opportunity: Theory and evidence for Germany and the US. *Social Choice and Welfare*, 43, 73-99.

Nygård, F., & Sandström, A. (1989). Income inequality measures based on sample surveys. *Journal of Econometrics*, 42(1), 81-95.

O'rand, A. & Henrettam, J.C. (2018). Age and inequality: Diverse pathways through later life. *Routledge.*

Organization for Economic Co-operation and Development, OECD (2008). Handbook on Constructing Composite Indicators. Methodology and user guide. *OECD Publishing.*

Organization for Economic Co-operation and Development , OECD (2016). OECD Regional Outlook 2016: Productive Regions for Inclusive Societies. Chapter 3: Understanding rural economies. OECD Publishing, Paris

Palomino, J.C., Marrero, G.A., & Rodríguez, J. A. (2019). Channels of inequality of opportunity: The role of education and occupation in Europe. *Social Indicators Research,* 143, 1045-1074.

Pandeya, B., Buytaert, W., Zulkafli, Z., Karpouzoglou, T., Mao, F., & Hannah, D. M. (2016). A comparative analysis of ecosystem services valuation approaches for application at the local scale and in data scarce regions. *Ecosystem Services*, 22, 250-259.

Pike, A., Béal, V., Cauchi-Duval, N., Franklin, R., Kinossian, N., Lang, T., & Velthuis, S. (2023). 'Left behind places': a geographical etymology. *Regional Studies*, 1-13.

Ramos, X., & Van de gaer, D. (2016). Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys*, 30(5), 855-883.

Ramos, X., & Van de gaer, D. (2021). Is inequality of opportunity robust to the measurement approach? *Review of Income and Wealth*, 67(1), 18-36.

Robeyns, I. (2017). Wellbeing, freedom and social justice: The capability approach re-examined. UK: *Open Book Publishers.*

Roch-Dupré, D., Gonsalves, T., Cucala, A.P., Pecharromán, R.R., López-López, A.J., Fernández Cardador, A. (2021). Determining the optimum installation of energy storage systems in railway electrical infrastructures by means of swarm and evolutionary optimization algorithms. *International Journal of Electrical Power & Energy Systems*, 124, 106295-1 - 106295-15

Roch-Dupré, D., Gonsalves, T., Cucala, A.P., Pecharromán, R.R., López-López, A.J., Fernández Cardador, A. (2021). Multi-stage optimization of the installation of energy storage systems in railway electrical infrastructures with nature-inspired optimization algorithms. *Engineering Applications of Artificial Intelligence*, 104, 104370-1 - 104370-18

Roemer, J. E. (1993). A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs*, 22 (2), 146-166.

Roemer, J. E. (2000). Equality of opportunity. *Harvard University Press.*

Roemer, J. E., & Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, 54(4), 1288-1332.

Royuela, V., P. Veneri & R. Ramos (2014). Income inequality, urban size and economic growth in OECD regions. *OECD Regional Development Working Papers* No. 2014/10.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Sanogo, T. (2019). Does fiscal decentralization enhance citizens' access to public services and reduce poverty? Evidence from Côte d'Ivoire municipalities in a conflict setting. *World Development*, 113, 204-221.

Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., Fuller, G. (2018). SDG Index and dashboards report 2018. *Bertelsmann Stiftung and Sustainable Development Solutions Network.*

Saunders, P. (1995). The immigrant dimension of income inequality. *Growth*, 43, 66-86.

Sen, A. (1999). Development as freedom. *Oxford University Press.*

Shek, D. T., & Wu, F. K. (2018). The social indicators movement: Progress, paradigms, puzzles, promise and potential research directions. *Social Indicators Research*, 135, 975-990.

Shin, K. S., & Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, 23(3), 321-328.

Silveira, R. D. M., & Azzoni, C. R. (2011). Non-spatial government policies and regional income inequality in Brazil. *Regional Studies*, 45(4), 453-461.

Soleimani H., Govindan, K., Saghafi, H., Jafari, H. (2017). Fuzzy multi-objective sustainable and green closed-loop supply chain network design. *Computers & Industrial Engineering*, 109, 191–203.

Somarriba, N., & Pena, B. (2009). Synthetic indicators of quality of life in Europe. *Social Indicators Research*, 94, 115-133.

Taylor, S.J. & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37-45.

Wilkinson, R. G., & Pickett, K. E. (2009). Income inequality and social dysfunction. *Annual Review of Sociology*, 35, 493-511.

World Bank (2005). World development report 2006: equity and development. *World Bank.*

Yerkes, M., Javornik, J. & Kurowska, A. (2019). Social policy and the capability approach: Concepts, measurements and application. *Bristol University Press.*

Yitzhaki, S., & Schechtman, E. (2013). The Gini methodology: A primer on a statistical methodology. *Springer.*

Appendix A: Genetic algorithm to build a synthetic indicator

The genetic algorithm selects a subset of recurrent indicators, X_R^*, , that best explain a benchmark indicator Y_{NR}. The iterative process is as follows.

For the selection of recurrent indicators from X_R that should be included in X_R^*, we begin by defining any j-th solution A_j as a binary vector of size N_R, where N_R is the total amount of recurrent indicators available, such that

$$A_j = \left(a_{j,1}, a_{j,2}, \dots a_{j,N_R}\right), \qquad \text{A.1}$$

where $a_{j,i}$ can take the values of 0 or 1 for all $i \in \{1, \dots, N_R\}$. Whenever $a_{j,i} = 0$, indicator $x_{R,i}$ is not included in subset $X_{R,j}$. Therefore, a possible solution A_j denotes the combination of indicators $X_{R,j}$ constituted only by those indicators $x_{R,i}$ for which $a_{j,i} = 1$.[10]

The algorithm is initiated by generating an amount of J possible solutions A_j for all $j \in \{1, \dots, J\}$ in a first iteration. Each A_j is generated randomly, such that the probability of $a_{j,1} = 1$ is $0.5 \; \forall \; i \in \{1, \dots, N_R\}$ and $\forall \; j \in \{1, \dots J\}$. We therefore have J random combinations A_j of recurrent indicators, each denoted by $X_{R,j}$.

The genetic algorithm then works through the search space of possible combinations A_j to find the combination that maximises the R^2-statistic in a regression of the benchmark indicator Y_{NR} against $X_{R,j}$ while complying with the nonnegative restriction. In particular, any combination A_j of indicators is assessed by performing the following OLS regression:

$$Y_{NR} = \alpha_j + \omega_{R,j} X_{R,j} + \varepsilon_j, \qquad \text{A.2}$$

[10] Each A_j conceptually represents a unique combination of recurrent indicators. Note that because each $a_{j,i}$ can take two values (0 and 1), the total number of possible combinations of high frequency indicators is 2^{N_R}. Therefore, the search space of combinations increases exponentially with a higher number of available indicators, which entails the use of an optimization algorithm.

where α_j and $\omega_{R,j}$ are the OLS constant and coefficients, respectively, and ε_j is a vector of error term. Next, a fitness value is assigned to A_j such that

$$Fitness(A_j) = \begin{cases} R_j^2 & if \quad \omega_{R,j} \geq 0; \\ R_j^2 - \lambda & otherwise, \end{cases} \qquad \text{A.3}$$

where $\omega_{R,j}$ corresponds to the OLS estimates of the weights in Equation A.2, R_j^2 is the estimated R^2-statistic, and λ is a penalisation factor. Whenever the nonnegative restriction is not complied with, such that any element in $\omega_{R,j} < 0$, the fitness value of A_j will be heavily penalised with a large value, λ. Because the genetic algorithm maximises $Fitness(A_j)$, it dismisses all combinations A_j of indicators that violate the nonnegative restriction while continuing to search for the combination that provides the highest R^2-statistic.

Once the fitness values are calculated for all the randomly generated combinations of indicators A_j, these values are rescaled into probabilities. To do so, the combinations A_j, for all $j \in \{1, \dots, J\}$, are first ranked from highest to lowest according to their fitness value. Next, the scaled probability p for each A_j is defined as

$$p(A_j) = \frac{1}{\sqrt{r_j}}, \qquad \text{A.4}$$

where r_j is the rank of individual A_j.

In a second iteration of the algorithm, a new set of J possible solutions A_j will be created from the previous set. This is performed by first randomly selecting combinations A_j according to their scaled probabilities $p(A_j)$. New combinations A_j will then be created by two specific functions of the genetic algorithm denoted *crossover* and *mutation*, which mimic the evolutionary theories put forward by Charles Darwin. With *crossover*, two of the randomly selected combinations A_j are blended. The *genetic algorithm uses the crossover function* to explore the search space of possible combinations of indicators in its task for optimisation. With *mutation*, one of the randomly selected combinations A_j is altered to provide diversity to the possible combinations A_j to avoid premature convergence to a solution.

Once the new set of J possible solutions is created, the process is repeated by calculating the fitness values of the new combinations, rescaling these fitness values into probabilities, and selecting combinations for *crossover* and *mutation*. This is iterated numerous times until the algorithm converges to an optimal solution, denoted A^*, and defined as

$$A^* = \left(a_1^*, a_2^*, \dots, a_{N_R}^*\right). \qquad \text{A.5}$$

Subset X_R^* will then include all indicators $x_{R,i}$ for which $a_i^* = 1, \forall\, i \in \{1, \dots, N_R\}$.

Appendix B: Recurrent indicators for circumstances underlying income inequality across municipalities in Madrid

Table B1 summarises the recurrent indicators that depict the circumstances underlying income inequality in the municipalities of Madrid.

Table B1

Categorisation of recurrent indicators underlying income inequality.

Category	Recurrent indicators on circumstances		Studies linking circumstances to income inequality
Demography	Total population	Dependency ratio	Aaberge & Brandolini (2015), Cabrera et al. (2021), De Barros et al. (2009), Marrero & Rodríguez (2012), O'rand & Henretta (2018)
	Female population	Foreign population	
	Youth population	Foreign female population	
	Senior population		
Labour market	Working ratio**	Unemployment relative variation	Marrero & Rodríguez (2012)
	Female working population**	Youth unemployment	
	Foreign working population**	Female youth unemployment	
	Young working population**	Foreigners' unemployment	
	Senior working population**	Female work insertion**	
	Temporary contracts	Foreign intra-EU work insertion**	
	Unemployment rate	Foreign extra-EU work insertion**	
	Female unemployment		
Income	GDP per capita**	Urban tax base per receipt**	Bourguignon et al. (2007), Hufe et al. (2018), Lefranc et al. (2008)
	Number of tax declarations**	Labour income**	
	Tax base amount**	Gross disposable income**	
	Taxable saving base**	Families with minimum insertion income*	
Living conditions	Electricity consumption**	Enrolment rate for basic education**	Brock (2020), Bouzarovski & Tirado-Herrero (2017), Chatterjee & Turnovsky (2012), Gamboa & Waltenberg (2012), Jusot et al. (2013), Kilkiş (2016), Marmot (2005), Niehues & Peichl (2014), OECD (2016)
	Sanitary infrastructure**	Students per teacher	
	Water consumption**	Students per school unit	
	Passenger cars**	Public education	
	Population dispersion **		

*Base indicators divided by the municipality population to eliminate the effect of the municipality size.
**Base indicators inverted to ensure their unambiguity.

The data are aggregated and made available by the Regional Statistics Office since 2009. Tables B2 to B5 depict each indicator's description and primary source across categories.

Table B2

Demography recurrent indicators.

Indicator	Description
Total Population	Total number of individuals living in a municipality - January 1st of a given year
Female Population	Total number of women living in each municipality over total population
Youth Population	Total number of individuals between 15 and 24 years old over total population
Senior Population	Total number of individuals aged 65 over total population
Dependency Ratio	Population aged < 15 + Population aged > 65 divided by the population between 15 and 64
Foreign Population	Total number of individuals with foreign nationality over total population
Foreign Female Population	Total number of women with foreign nationality over total foreign population

Primary Source: Economic Management Directory. Madrid's Regional Government Data retrieved from the National Statistics Office.

Table B3

Labour Market recurrent indicators.

Indicator	Description
Working Ratio[1]	Number of people employed over the total population of working age
Female Working Population[1]	Total number of female workers over total working population
Foreign Working Population[1]	Total number of foreign workers with foreign nationality over total working population
Young Working Population[1]	Total number of individuals between 15 and 24 years working over total working population
Senior Working Population[1]	Total number of individuals aged > 65 working over total working population
Temporary Contracts[1]	Total number of temporary contracts over total working population
Unemployment Rate[2]	% of total labour force that is unemployed but actively seeking employment
Female unemployment[2]	Total number of women individuals unemployed and actively seeking employment
Unemployment relative variation[2]	% change in registered unemployment on March 31st in year $(t-1)$ and March 31st in year (t)
Youth unemployment[2]	Total number of individuals between 15 and 24 years old unemployed and actively seeking employment over total unemployment
Female youth unemployment[2]	Total number of women between 15 and 24 years old unemployed and actively seeking employment over total unemployment
Foreigners' Unemployment[2]	Total number of foreign individuals unemployed and actively seeking employment over total unemployment
Female Work Insertion[2]	Total new registered contracts for women over total new contracts
Foreign Intra-EU Work Insertion[2]	Total new registered contracts for EU nationals over total new contracts
Foreign Extra-EU Work Insertion[2]	Total new registered contracts for non-EU nationals over total new contracts

[1] Primary Source: Registered Contracts Statistics. Labour and Social Economy Ministry.

[2] Primary Source: Employment Statistics. Labour and Social Economy Ministry.

Table B4

Income recurrent indicators.

Indicator	Description
GDP per Capita[1]	Gross domestic product per individual within each municipality
Number of Tax Declarations[2]	Total number of tax declarations filed in each municipality per capita
Tax base amount[2]	Personal per capita income tax - Tax base amount
Taxable saving base[2]	Personal per capita income tax - Taxable savings base amount
Urban tax base per receipt[2]	Amount of assets or income that can be taxed by a municipality per receipt
Labour income[3]	Total income from labour-related sources per capita
Gross disposable income[3]	Income available to households for consumption or investment, after redistribution operations.
Families with Minimum Insertion Income[4]	Number of households receiving regional minimum income benefits

[1] Primary Source: Municipal GDP Indicators. Madrid's Regional Statistics Office.

[2] Primary Source: National Tax Office - Register.

[3] Primary Source: Municipal Household Income Indicators. Madrid's Regional Statistics Office.

[4] Primary Source: Social Security Statistics.

Table B5

Living Conditions recurrent indicators.

Indicator	Description
Electricity Consumption (1)	Average consumption of electricity per household
Sanitary infrastructure (2)	Local health centers per 10.000 residents
Water Consumption (3)	Average consumption of water consumed by each household
Passenger cars (4)	Total number of passenger cars per 1.000 individuals
Population Dispersion (5)	Average population per square kilometer of a municipality
Enrollment rate for basic education (6)	Ratio of pupils, students and apprentices enrolled in basic education over total youth population
Students per teacher (6)	Ratio of pupils, students, and apprentices per teacher
Students per School Unit (6)	Total number of students divided by total number of schools within a municipality
Public education (6)	Percentage of non-university students enrolled in public schools

1 Primary Source: Private companies.

2 Primary Source: Madrid Health Service. Department of Health Madrid's Regional Government.

3 Primary Source: Public Company - Canal Isabel II.

4 Primary Source: Transit Authority. Ministry of Domestic Affairs.

5 Primary Source: Economic Management Directory. Madrid's Regional Government Data.

6 Primary Source: Sub directorate General for Evaluation and Analysis. Department of Education, Universities, Science and Spokesperson Madrid's Regional Government.

Appendix C. Descriptive Statistics

	Mean	Median	Mode	Std. Dev.	Variance	Skewness	Kurtosis
80/20 Poverty Ratio	2,995	2,9	2,7	0,417	0,174	1,243	6,792
GINI	33,594	33,3	35,7	3,098	9,597	0,393	2,931
Demography							
Total Population	0,09	0,017	1	0,192	0,037	3,186	13,133
Female Population	0,925	0,94	1	0,059	0,004	-2,237	9,057
Youth Population	0,617	0,628	1	0,157	0,025	-0,602	4,112
Senior Population	0,354	0,336	0,243	0,146	0,021	1,234	5,526
Dependency Ratio	0,491	0,479	0,5	0,108	0,012	1,404	7,375
Foreign Population	0,432	0,413	1	0,178	0,032	0,451	3,105
Foreign Female Population	0,554	0,526	0,5	0,11	0,012	-0,222	7,28
Labour Market							
Working Ratio	0,388	0,375	1	0,166	0,028	0,688	3,84
Female Working Population	0,627	0,614	0,582	0,069	0,005	2,118	10,798
Foreign Working Population	0,211	0,171	0,081	0,143	0,02	2,908	13,434
Young Working Population	0,203	0,185	0,138	0,089	0,008	5,32	42,367
Senior Working Population	0,669	0,672	0,483	0,128	0,016	-0,145	2,91

Temporary Contracts	0,571	0,578	0,389	0,149	0,022	0,137	2,11
Unemployment Rate	0,513	0,498	0,554	0,168	0,028	0,386	2,985
Female unemployment	0,541	0,548	0,5	0,095	0,009	-0,555	13,152
Unemployment relative variation	-0,14	-0,136	-0,25	0,201	0,04	0,189	14,982
Youth unemployment	0,21	0,155	0,078	0,165	0,027	1,386	5,902
Female youth unemployment	0,443	0,455	0	0,205	0,042	0,022	4,467
Foreigners' Unemployment	0,373	0,344	0,286	0,177	0,031	0,711	3,733
Female Work Insertion	0,171	0,154	0,171	0,093	0,009	5,764	47,984
Foreign Intra-EU Work Insertion	0,199	0,169	0,181	0,137	0,019	2,483	12,315
Foreign Extra-EU Work Insertion	0,21	0,163	0,155	0,156	0,024	2,407	10,157
Income							
GDP per Capita	0,417	0,403	1	0,17	0,029	0,434	3,207
Number of Tax Declarations	0,424	0,405	1	0,077	0,006	3,287	21,514
Tax base amount	0,333	0,311	1	0,135	0,018	1,477	7,11
Taxable saving base	0,374	0,401	1	0,14	0,02	0,722	5,099
Urban tax base per receipt	0,123	0,088	1	0,132	0,018	3,757	21,708
Labour income	0,685	0,693	1	0,134	0,018	-0,408	3,107
Gross disposable income	0,242	0,19	1	0,171	0,029	1,387	5,095

Families with Minimum Insertion Income	0,213	0,152	1	0,182	0,033	1,636	5,892
Living Conditions							
Electricity Consumption	0,445	0,438	1	0,168	0,028	0,169	3,098
Sanitary infrastructure	0,039	0,008	0,417	0,111	0,012	6,076	47,027
Water Consumption	0,008	0,002	1	0,076	0,006	12,987	170,431
Passenger cars	0,689	0,74	0,783	0,211	0,044	-1,866	6,213
Population Dispersion	0,049	0,013	1	0,106	0,011	5,234	40,462
Enrollment rate for basic education	0,05	0,029	0	0,108	0,012	5,376	39,233
Students per teacher	0,607	0,725	0	0,314	0,098	-0,88	2,44
Students per School Unit	0,623	0,735	0	0,318	0,101	-0,89	2,49
Public education	0,718	0,936	1	0,356	0,127	-1,037	2,647