Socio-Economic Planning Sciences 91 (2024) 101784

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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

# An attempt to correct the underestimation of inequality measures in cross-survey imputation through generalized additive models for location, scale and shape

Gianni Betti [a], Vasco Molini [b], Lorenzo Mori [c,*]

[a] The University of Siena Department of Economics and Statistics, Italy
[b] 50x2030 Initiative, Italy
[c] The University of Bologna, Department of Statistical Sciences "Paolo Fortunati", Italy

ARTICLE INFO

*Keywords:*
Bias reduction
Inequality indicators
Moroccan HBS
Moroccan LFS
Survey-to-survey imputation

ABSTRACT

This paper contributes to the debate on ways to improve the calculation of inequality measures in developing countries experiencing severe budget constraints. Linear regression-based survey-to-survey imputation techniques (SSITs) are most frequently discussed in the literature. These are effective at estimating predictions of poverty indicators but are much less accurate with inequality indicators. To demonstrate this limited accuracy, the first part of the paper review and discuss the SSITs. The paper proposes a method for overcoming these limitations based on a Generalized Additive Models for Location, Scale and Shape (GAMLSS). Before to apply this method to Moroccan data with the aim to analyze the relation between poverty and climate changes a simulation is carried out to compare classical SSIT and SSIT based on GAMLSS.

## 1. Introduction

The estimation of poverty and inequality measures is a primary goal for every National Statistical Offices (NSOs). Typically, these estimates arise from survey's data collected annually in order, not only to estimate such indicators, but also to study and analyze possible scenarios and trends from on year to another. The collection of these data requires elaborate and expensive surveys about consumption expenditure (such as Household Budget Surveys, HBS) or about income (as the EU statistics on income and living conditions, EU-SILC). Only few countries can collect data annually to facilitate the estimation of poverty and inequality. Therefore, producing reliable indices annually for monitoring poverty results remains quite challenging for many countries. To overcome this challenge, scholars have focused on developing methods to compare welfare indicators over time from surveys that are little comparable.

These techniques, broadly known as Survey-to-Survey Imputation Techniques (SSITs), proved successful at predicting comparable poverty indicators, but as this paper argues, were less effective at predicting comparable inequality indicators. SSITs techniques come from the poverty map field [1,2]. Poverty map is a technique that aims to produce reliable estimates imputing income into census. Nowadays, this method was also used for SSITs mapping from surveys with consumption data to those with other outcomes of interest [3–5].

While this approach is well established for poverty estimates, a cursory overview of the literature shows that little attention is devoted to potential problems in obtaining accurate inequality measures. In the following we report the most common limitation of SSITs highlighted by authors. Demombynes and Hoogeveen [6] using data from rural Mexican communities show as the SSITs estimates values, when compared with the true ones, are more correlated with poverty measures than with inequality measures. Newhouse et al. [7] argue that Survey-to-Survey Imputation can fail. They demonstrate that minor differences in the sampling scheme, sampling design, or structure of the answers/questions can produce inaccurate SSITs. Douidich et al. [8] are able to obtain accurate estimates for quarterly poverty rates using a log-linear regression to impute the total expenditure using an exogenous poverty line but no information are given in the case of an endogenous poverty line. In their 2019 study, Krafft et al. [26] impute consumption expenditure from Household Budget Surveys (HBSs) to Labor Force Surveys (LFSs) to investigate poverty and inequality in Jordan and the Arabic Republic of Egypt. They aim to obtain inequality estimates for those years where only the LFS, which originally has not the consumption variable, is available. The model used to impute the values is a

* Corresponding author.
*E-mail addresses:* gianni.betti@unisi.it (G. Betti), vmolini@worldbank.org (V. Molini), lorenzo.mori7@unibo.it (L. Mori).

https://doi.org/10.1016/j.seps.2023.101784
Received 12 August 2023; Received in revised form 29 November 2023; Accepted 6 December 2023
Available online 13 December 2023

classic log-normal model which is time invariant with normal and homoscedastic error term. The findings reveal comparable levels of consumption expenditure, poverty, and inequality between the surveyed pairs. However, the concern is that incorrectly assuming normal errors or ignoring heteroscedasticity has the potential to introduce bias when estimating poverty or inequality. Specifically, in the prediction of inequality, the bias can be significant. Given that the SSITs are a derivative of this method, we argue their conclusions can also apply to SSITs.

We argue that SSITs, by construction, tend to underestimate the predicted inequality, and to obtain accurate estimates, a different methodological approach should be chosen. The assumption of residuals' normality distribution and the fact that standard SSITs are based on linear regression make them more accurate at predicting the central moments of a distribution (or transformations such as the poverty headcount) rather than the shape of the tails. This latter is crucial in predicting inequality. These models, however, tend to predict distribution compressed around the mean and, with thin tails, underestimate inequality. This point is not critical if the outcome parameters are related with poverty, however, as shown by Schluter [9], it is a crucial aspect of inequality measures. Moreover, the linear regression in SSITs is often reached thanks to the Box-Cox transformation (or logarithmic transformation) the use of the transformation introduces the problem of the back transformation as well [10].

To overcome these limitations, we propose to use to substitute the linear regression with Generalized Additive Models for Location, Scale and Shape (GAMLSS, [11]). GAMLSS, in our view, can be easily extent to the SSITs theory and can improve estimate mainly for three reasons. First of all, GAMSS completely release the exponential family distribution assumption, i.e. there is not the necessity to assume the normality of the data, allowing researchers to use more than 100 different probabilistic distribution. Secondly, GAMLSS allows to use covariates not only on the location parameter (as linear regression) but to define each parameter of the distribution in terms of covariates. Thirdly, the time invariance can be easily overcome. Summing up, GAMLSS could increase the fitting of the distribution through the possibility to define the distribution that best fit the data and using covariates on all the parameters of the distribution and can easily delete the time invariance which is often assumed in SSITs.

Moreover, it is important to note two additional things which could be important. Most of the distribution have a closed expression for the inequality parameters [12,13]. Rarely they depend only on the location parameter and hence the possibility to use covariates also on scale and shape parameters could be fundamental. Secondly, GAMLSS allow the users to use incorporate in the regression also an "Additional" term which, in this specific case, can be a random effect that can help to highlight difference between groups/areas.

In general, these new techniques could be seen as an extension of the classical SSITs in which the normal distribution assumption is replaced with the distribution that fit as better as possible the data and where the fixed effects used in SSITs are replace with random effects. We tested this approach on a dataset recently used in Morocco [8].

The paper is organized as follows: section 2 overview SSITs and their problems; section 3 presents the new SSIT-GAMLSS techniques; section 4 presents the data used in section 5 for the application and in section 6 for simulations. Section 7 concludes with some remarks and possible extension of the work.

## 2. A critical overview of survey-to-survey regression techniques

In general, SSITs use a log-linear regression to estimate values for the regression parameter at year *t*, then use those values to predict the dependent variable for the previous years where only the independent variables are available. This process is done under the hypothesis that the parameters are stable over the years. Often, this hypothesis is verified by estimating parameters at two non-consecutive years to illustrate that they are similar irrespective of the selected years. The regression usually uses individual or household expenditures as the dependent variable. With the right skew and in order to avoid normality, the variable is transformed into a logarithm and then transformed back to obtain the predicted values. In addition, to mitigate the problem of the underestimation of the teil the predicted values are summed with an error component as usual for the techniques based on the poverty mapping [1].

Regression is the best way to obtain predicted values only if all the assumptions are fulfilled and the selected dependent variables have a higher explanatory power versus the independent one. This last point is summarized in $R^2_{adj}$. Often, articles do not report the $R^2_{adj}$, although multiple regressions are always used. Interaction or special transformation of the dependent variables is also common. Unfortunately, failure to report the $R^2_{adj}$ hinders the ability to check the accuracy of the regression. Given the connection between $R^2_{adj}$ and $R^2$ we can argue that the first is at least equal to or—more likely—lower than the ones reported (which is often around 0.5 or less). A low $R^2$ can result from a selection that prefers a common set of variables from the two surveys—which is a key requirement for SSITs—over a high explanatory power [5,8,14].

Demographic variables are always used in those type of analysis but, as highlighted by Ketkar et al. [15], household sociodemographic characteristics are as important as determinants of expenditure patterns as price and income. The standard consumer behavior model states that the economic agent maximizes her utility subject only to relative prices and to income constraints. The two theories may both be valid, but if prices and income are excluded from the standard SSITs models, the sociodemographic variables alone may not adequately explain the individual or household consumption patterns. However, these conclusions must be verified for every dataset and are impossible to determine a priori.

A low value for $R^2$ arising from insufficient correlated independent variables, combined with a logarithmic transformation, could also generate problems with establishing predictions. The logarithmic transformation compresses the tail of skewed and kurtotic variables, which effectively generate symmetric PDFs and, therefore, cause Gaussian-like errors. But the predicted values could be applied in the tails and, given the low $R^2_{adj}$, they could all be near the mean, with a sharp decrease in data variability. In addition, logarithmic transformations are not immune to a back-transformation bias [10]. There are various ways to reduce the back-transformation bias, but the most frequently used ones are based on a scale correction of the predicted values. Unfortunately, given that the Gini index and the Theil index are scale invariant,[1] those type of corrections are ineffective at reducing the bias.

To summarize the discussion thus far, the regression to obtain a predictor is the right choice only if the $R^2_{adj}$ is high, all the assumptions are respected, and the bias caused by the log-transformation is negligible. In addition, it is important to bear in mind other potential sources of bias [7] arising from the differences between the survey design and the questionnaires.

## 3. The proposal: A SSIT based on generalized additive models for location scale and shape

Proposed by Rigby and Stasinopoulos [11], GAMLSS incorporates location parameter, scale parameter, and shape parameters. These

[1] The property of scale invariance states that inequality remains unchanged when all incomes increase by the same proportion. See Clementi et al. [25] for a discussion of differences between (relative) scale invariant and non-scale invariant (absolute) measures of inequality.

models expand the hypothetical distribution form beyond the exponential distribution family, encompassing a wide range of commonly encountered distribution types.

GAMLSS assume independent observations $y_i, i=1,\ldots,n$ from a random variable $Y$, with Probability Density Function (PDF) $f(\mathrm{Y}|\theta_i)$, conditional on a vector of $p$ distribution parameters, $k=1,\ldots,p$ ($\boldsymbol{\theta}_i^T = (\boldsymbol{\theta_{i1}},\ldots,\boldsymbol{\theta_{ik}},\ldots,\boldsymbol{\theta_{ip}})$). More formally, let $\mathbf{y}^T=(y_1,\ldots y_n)$ be the $n$ length vector of the response variable. Let $g_k(\cdot)$ be a known monotonic link functions relating the $p$ distribution parameters to explanatory variables by:

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}^k\boldsymbol{\beta}_k + \sum_{m=1}^{M_k} \mathbf{Z}_m^k\gamma_m^k, \text{with } k=1,\ldots,p \tag{1}$$

where $\boldsymbol{\theta}_k^T=(\boldsymbol{\theta_{1k}},\ldots,\boldsymbol{\theta_{nk}})$ is a vector of length $n, \boldsymbol{\beta}_k^T=(\boldsymbol{\beta_{1k}},\ldots,\boldsymbol{\beta_{kM_k}})$ is a parameter vector of length $M_k, \mathbf{X}^k$ is a matrix of known covariates of order $n\times M_k, \mathbf{Z}_m^k$ is a fixed known $n\times q_{mk}$ design matrix and $\boldsymbol{\gamma_m^k}$ is a $q_{mk}$-dimensional random variable. A number of different additive smoothing terms are allowed in (1). Changing the definition of the matrix $\mathbf{Z}_m^k$ is possible to include P-spline, cubic splines, random-effects, non-parametric random effects, and many others.

### 3.1. Adding geographical component

SSITs are typically employed to generate reliable estimates either at the national level or at a more detailed disaggregated level, aligning with the surveys' intended purpose. SSITs predominantly incorporate fixed effects. In the subsequent discussion, we suggest employing a variation of specification (1) that incorporates random effects. The idea is to use a specific random effect based on the planned disaggregated level according with the surveys. Let us introduce the random effects more formally.

We identify with $\boldsymbol{Y_{ij}}$ the target variable in unit $i$ from area $j$, with $i=1,\ldots,n$ and $j=1,\ldots,J$. In this context, the aim is to estimate area parameters in the form of $\boldsymbol{H}_j=\zeta(\boldsymbol{Y}_{ij}), j=1,\ldots,J$, where $\zeta(\cdot)$ is a real measurable function. Moving from (1) we propose a SSIT-GAMLSS by considering area specific random effect and limiting our attention to four or less parameter distributions ($k=1,\ldots,4$):

$$\begin{cases} g_\mu(\boldsymbol{\mu}_{ij}) = \mathbf{X}_{ij}^\mu\boldsymbol{\beta}_\mu + \boldsymbol{\gamma}_j^\mu \\ g_\sigma(\boldsymbol{\sigma}_{ij}) = \mathbf{X}_{ij}^\sigma\boldsymbol{\beta}_\sigma + \boldsymbol{\gamma}_j^\sigma \\ g_\nu(\boldsymbol{\nu}_{ij}) = \mathbf{X}_{ij}^\nu\boldsymbol{\beta}_\nu + \boldsymbol{\gamma}_j^\nu \\ g_\tau(\boldsymbol{\tau}_{ij}) = \mathbf{X}_{ij}^\tau\boldsymbol{\beta}_\tau + \boldsymbol{\gamma}_j^\tau \end{cases} \tag{2}$$

Note as (2) is equivalent to (1) with $\mathbf{Z}=\mathbb{I}_\gamma$ and $M_k=1$. In (2) random effects are $\boldsymbol{\gamma}_j^k \overset{\sim}{iid} N(\mathbf{0},\Psi_k)$ for $k=1,\ldots,4$. The variance-covariance matrix $\Psi$ of the multivariate Normal involves the variance of the random effects $\boldsymbol{\sigma}_k^2$.

We use the function fitDist() in the package GAMLSS [16] to fit all relevant parametric distributions to a single data vector to choose the distribution that fit our data at the best. The final marginal distribution is selected by the Generalized Akaike Information Criterion (GAIC). Following Rigby et al. [17], we use a penalization equal to $\sqrt{\log(n)}\simeq 2$ where $n$ is total number of sampled units. Once that the best distribution is selected (2) a penalized log-likelihood is used to obtain the estimates of the regression parameters. The quadratic penalties in the likelihood result from assuming a normally distributed random-effect on the linear predictor. The chosen distribution is denoted as $\mathscr{T}(\boldsymbol{\mu}_{ij},\boldsymbol{\sigma}_{ij},\boldsymbol{\nu}_{ij},\boldsymbol{\tau}_{ij})$.

### 3.2. From one-year estimates to SSIT-GAMLSS

As mentioned earlier, SSITs are typically time-invariant. That is to say that regression parameters are derived from the most recent dataset and are subsequently applied to generate imputed values for the preceding year, irrespective of the number of years that have elapsed.

Let us now introduce the time components and add this to our dependent variable: $\boldsymbol{y_{ij}}^t$. In other words, $\boldsymbol{y_{ij}}^t$ is our dependent variable observed at time $t$ for unit $i$ in area $j$. As noted, so far, we have to estimate a quantity $\boldsymbol{H}_j^t=\zeta(\boldsymbol{Y}_{ij}^t)$ from year $t_-$ to year $t_+$, with $t_-<t_+$, and having the dependent variable only at $t_-$ and at $t_+$. We propose, to estimate a GAMLSS at time $t_-$ and a GAMLSS at time $t_+$ and to use a weighted mean of the estimated parameters between $t_-$ and $t_+$. Let us assume independence between estimates at time $t_-$ and at time $t_+$ and denote with the super-script $t_-$ or $t_+$ the estimated parameters, i.e. $\beta_k^{t_-}, \boldsymbol{\gamma}_j^{kt_-}$ and $\beta_k^{t_+}, \boldsymbol{\gamma}_j^{kt_+}\forall k$. At a generic time $t^*: t_-\le t^*\le t_+$ we use as prediction model a GAMLLS based on a probabilistic distribution model with the following parameters:

$$\begin{cases} g_\mu\left(\boldsymbol{\mu}_{ij}^{t^*}\right) = \mathbf{X}_{ij}^{\mu t^*}\boldsymbol{\beta}_\mu^{t^*} + \boldsymbol{\gamma}_j^{\mu t^*} = \mathbf{X}_{ij}^{\mu t^*}\left(\omega\boldsymbol{\beta}_\mu^{t_-} + (1-\omega)\boldsymbol{\beta}_\mu^{t_+}\right) + \left(\omega\boldsymbol{\gamma}_j^{\mu t_-} + (1-\omega)\boldsymbol{\gamma}_j^{\mu t_+}\right) \\ g_\sigma\left(\boldsymbol{\sigma}_{ij}^{t^*}\right) = \mathbf{X}_{ij}^{\sigma t^*}\boldsymbol{\beta}_\sigma^{t^*} + \boldsymbol{\gamma}_j^{\sigma t^*} = \mathbf{X}_{ij}^{\sigma t^*}\left(\omega\boldsymbol{\beta}_\sigma^{t_-} + (1-\omega)\boldsymbol{\beta}_\sigma^{t_+}\right) + \left(\omega\boldsymbol{\gamma}_j^{\sigma t_-} + (1-\omega)\boldsymbol{\gamma}_j^{\sigma t_+}\right) \\ g_\nu\left(\boldsymbol{\nu}_{ij}^{t^*}\right) = \mathbf{X}_{ij}^{\nu t^*}\boldsymbol{\beta}_\nu^{t^*} + \boldsymbol{\gamma}_j^{\nu t^*} = \mathbf{X}_{ij}^{\nu t^*}\left(\omega\boldsymbol{\beta}_\nu^{t_-} + (1-\omega)\boldsymbol{\beta}_\nu^{t_+}\right) + \left(\omega\boldsymbol{\gamma}_j^{\nu t_-} + (1-\omega)\boldsymbol{\gamma}_j^{\nu t_+}\right) \\ g_\tau\left(\boldsymbol{\tau}_{ij}^{t^*}\right) = \mathbf{X}_{ij}^{\tau t^*}\boldsymbol{\beta}_\tau^{t^*} + \boldsymbol{\gamma}_j^{\tau t^*} = \mathbf{X}_{ij}^{\tau t^*}\left(\omega\boldsymbol{\beta}_\tau^{t_-} + (1-\omega)\boldsymbol{\beta}_\tau^{t_+}\right) + \left(\omega\boldsymbol{\gamma}_j^{\tau t_-} + (1-\omega)\boldsymbol{\gamma}_j^{\tau t_+}\right) \end{cases} \tag{3}$$

where $\omega=\frac{t_+-t^*}{t_+-t_-}$ with $t_-\le t^*\le t_+$ and the covariates matrices $\mathbf{X}_{ij}^{kt^*}, \forall k$ come, year by year, from a second surveys as usual in SSITs. Some remarks are necessary. The classical SSITs are a special case of the SSIT-GAMLSS. Assuming normality without the use of the random-effect and with $\omega=0$ for every $t^*$ they are exactly the same. Moreover, the definition of the weight $\omega$ leads to use the estimated model (2) at $t_-$ and $t_+$ given that $\omega$ will be equal to 1 and 0, respectively. To conclude as noted before random effects are assumed to be normal with 0 mean and variance equal to $\sigma_k^2, \forall k$ the new random effects which, basically, are a weighted mean of the original one estimated at $t_-$ and $t_+$ remain normal with 0 mean and variance equal to the weighted mean of the variance.

Lastly the prediction of the area parameter $\boldsymbol{H}_j^t=\zeta(\boldsymbol{Y}_{ij}^t)$ is obtained with a Monte-Carlo approach as follows:

1 fit the model (2) to the sample data, obtaining a consistent estimate of the model at time $t_-$ and $t_+$ ad generate the model (3) at every $t^*: t_-\le t^*\le t_+$;
2 for each $\ell=1,\ldots,L$, for $L$ large, generate the vector of imputed values ${}^{(\ell)}\mathbf{y}_{ij}^{\mathbf{t}}$;
3 Compute the target parameter $\boldsymbol{H}_j^{(\ell)\mathrm{t}}=\zeta({}^{(\ell)}\mathbf{y}_{ij}^{\mathbf{t}})$ for each $t$ using the sample weights;
4 A MC approximation of $\widehat{\boldsymbol{H}}_j^t$ is then:

$$\widehat{\boldsymbol{H}}_j^t \approx \frac{1}{L}\sum_{i=1}^{L}\boldsymbol{H}_j^{(\ell)\mathrm{t}} \tag{4}$$

The Mean Square Error (MSE) of estimates, following Rust and Rao [18] and Field and Welsh [19] is computed with classical non-parametric bootstrap for grouped data.

## 4. Data

In this section we present the two sets of surveys that were used. The HBS and the LFS of Morocco. The HBS is carried out every seven years, while the LFS is conducted yearly. The HBS reports total family expenditure, which we convert into per capita expenditure expressed in United States dollar. In Morocco this survey is made up of the 2000–2001 National Survey on Consumption and Expenditure (NSCE) and the 2006–2007 National Living Standards Survey (NLSS). NSCE and

**Table 1**
GAIC of distributions for equivalized consumption.

| Distribution | GAIC criterion (2000) | GAIC criterion (2007) |
|---|---|---|
| GB2 | 242586 | 122288 |
| Skew-t | 242707 | 122382 |
| Log-Normal | 243119 | 122678 |
| Pareto | 248864 | 125622 |
| Normal | 265796 | 133481 |

NLSS provide information about household expenditure and are representative up to urban and rural areas.

In the NSCE 15.000 households were sampled between November and October 2000. NLSS was smaller by sampling just 7.200 households and was conducted between December 2006 and November 2007. NSCE and NLSS have several sections in common however the former differentiates from the latter by including modules on transfers, subjective indicators of well-being, nutrition, and measurement for access to services. Among the shared sections, the most interesting from this work's point of view are the ones on socio-demographic characteristics, habitat, expenditures, durable goods, education, health, and employment.

NSCE and NLSS not only shared some modules but also the structure. They are both stratified at regional, provincial, and city (taking into account the size of the city between large, medium, and small) levels for urban areas. In this case, the sample includes five kinds of housing. Regarding rural areas, they are stratified only at the regional and provincial levels. These two surveys, however, share nothing but some modules and structure.

NSCE sample is obtained with a two-stage sampling scheme where, at the first stage, a list coming from the 1994 population census is used to extract the 300 households. For the NLSS the sampling scheme was changed and was added a third stage. Moreover, the list from which units are sampled come from the 2004 census and not from the 1994 one as for the NSCE. For the sake of completeness, it should be noted that the change of the list is not specific to this survey but that it has occurred for every survey carried out in Morocco since 2005.

To conclude, the LFS, which was first launched in 1976, follows a sampling process that is similar to that of the 2007 NLSS. Again, also for the LFS, the list from which the units are sampled was changed between 2005 and 2006 by introducing the one derived from the 2004 census. The LFS questionnaires share only a few questions with the other two surveys, relating to socio-demographic characteristics, habitat, education, health, and employment. LFS do not report any information about expenditure and/or monetary variables. What is important to note here, bearing in mind what asserted by Newhouse et al. [7], is that not only the sampling scheme change from a survey to another but also that the list from which the units are sampled come from two different census. In the following we will refer to both NSCE and NLSS as HBS.

## 5. Estimating inequality measures for Morocco over eight years

The main aim of this section is reporting and summarizing estimates from 2000 to 2007 for the Moroccan regions of three different inequality indicators: Gini index, Theil index and Atkinson index. The first step, as described in section 3, is the choice of the distribution according to the GAIC criterion. GAMLSS allows the use of over 100 of different distributions. In the following (Table 1) we report the GAIC of a number of selected distributions for both 2000 and 2007.

The smallest GAIC value, for both years, is reached by the four-parameters GB2 distribution, with log-link function for each parameter. The PDF of the GB2 is:

$$f_y(y|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu},\boldsymbol{\tau}) = \sigma y^{\sigma-1}\{\boldsymbol{\mu}^{\sigma\nu}B(\boldsymbol{\nu},\boldsymbol{\tau})[1+(y/\boldsymbol{\mu})^{\sigma}]^{\nu+\tau}\}^{-1} = \\ = \frac{\Gamma(\boldsymbol{\nu}+\boldsymbol{\tau})\boldsymbol{\sigma}(y/\boldsymbol{\mu})^{\sigma\nu}}{\Gamma(\boldsymbol{\nu})\Gamma(\boldsymbol{\tau})y[1+(y/\boldsymbol{\mu})^{\sigma}]^{\nu+\tau}} \quad (5)$$

**Table 2**
Regression coefficients estimates.

| Covariates | 2000 | 2007 |
|---|---|---|
| $\boldsymbol{\mu}$ (log-link function) | | |
| Intercept | 5.229*** | 6.825*** |
| Age | −0.013** | −0.012** |
| Sex | −0.716′ | −0.479* |
| Roompc | 3.049*** | 1.685*** |
| Primary | −1.097*** | −0.258*** |
| Secondary | −0.549 | −0.047 |
| Inactive | 0.326 | 0.032′ |
| $\boldsymbol{\sigma}$ (log-link function) | | |
| Intercept | −0.490*** | 0.363*** |
| Age | −0.001*** | −0.003* |
| Sex | −0.089*** | −0.189** |
| Roompc | −0.099*** | −0.133*** |
| Primary | −0.051* | −0.170* |
| Secondary | −0.077 | 0.172** |
| Inactive | 0.019 | 0.115* |
| $\boldsymbol{\nu}$ (log-link function) | | |
| Intercept | 3.937*** | 1.891*** |
| Age | 0.006** | 0.010** |
| Sex | 0.443** | 0.548** |
| Roompc | −0.823*** | −0.501*** |
| Primary | 0.346** | 0.410′ |
| Secondary | −0.134 | −0.337* |
| Inactive | −0.412** | −0.444** |
| $\boldsymbol{\tau}$ (log-link function) | | |
| Intercept | 2.853*** | 1.262*** |
| Age | 0.002* | 0.003 |
| Sex | 0.215*** | 0.298** |
| Roompc | 0.261*** | 0.305*** |
| Primary | −0.135** | 0.379** |
| Secondary | −0.662*** | −0.773*** |
| Inactive | −0.127* | −0.353** |
| $R^2_{adj}$ | 0.449 | 0.414 |

Signf. Codes: 0 *** 0.001 ** 0.01 * 0.05 ‘

for $y>0$, where $\boldsymbol{\mu}>0$,$\sigma>0$,$\boldsymbol{\nu}>0$ and $\boldsymbol{\tau}>0$ and where $B(\cdot)$ and $\Gamma(\cdot)$ are the Beta and Gamma function, respectively. In conclusion, the MC to estimate $\boldsymbol{H}_j^t = \zeta(\boldsymbol{Y}_{ij}^t)$ involves 200 iterations, and the bootstrap algorithm for estimating the MSE also relies on 200 iterations.

Before to move to the results let us summarize the covariates used for this analysis. We use age and sex of the householder adding “roompc” (room per capita) and the following dichotomous variable: “inactive” (the householder is an inactive person), “primary” (the householder is employed in the primary sector) and “secondary” (the householder is employed in the secondary sector). The estimated regression coefficients are reported in Table 2. Moreover, we use random effects (based on the urban/rural areas) on the location and scale parameters. We do not use random effects on the shape parameters because their variances are not statistically different from 0. The corresponding p-value is higher than 0.1. Table 3 summarize the variance of the random effects. A first important consideration to be done is the coherence of the estimated regression coefficients between 2000 and 2007 almost all the regression parameters are always statistically different from 0 and the variances of the random effects are significant. Some remarks are necessary. Let us start from the interpretability of the model. All the distribution parameters have a log-link function that leads to two possible ways to interpret the estimated coefficients: in their logarithmic form $log(\beta)<0$ $(log(\beta)>0)$ or equivalently in the exponential form $\beta<1$ $(\beta>1)$. In addition, the easiest way to interpret those type of regression, given that there is more than one distribution parameter, is verifying the impact of the estimated regression coefficient on the mean of the distribution. More precisely, the mean of a GB2 distribution is a function of $\mu$,$B(\nu+\sigma^{-1},\tau-\sigma^{-1})$ and $B(\nu,\tau)^{-1}$. When the householder is a female, keeping fixed the other regression coefficients, the values of $\mu$ decrease, $B(\nu+\sigma^{-1},\tau-\sigma^{-1})$ and $B(\nu,\tau)^{-1}$ decrease leading to a decrease also in the mean value of the dependent variables. In other words, the mean expenditure is lower if the householder is a female. Similarly, using a

**Table 3**
Estimated variance of random effects.

| Variance | 2000 | 2007 |
|---|---|---|
| $\mu$ | 0.338*** | 0.192*** |
| $\sigma$ | 0.029* | 0.028** |

Signf. Codes: 0 *** 0.001 ** 0.01 * 0.05 ‘

similar line of reasoning, we can conclude that the mean expenditure increases if the room per capita value increase. The variable roompc is also interesting if compared with the model developed by Douidich et al. [8]. In this study authors employed a log-linear regression model, creating two separate regressions for rural and urban areas. Our model shares only one covariate with theirs, specifically roompc. In our model, the remaining variables are either excluded or defined differently. In the Douidich et al. [8] model, roompc yields a positive regression coefficient that is statistically significant. Despite the substantial differences between the two models both arrive at the same conclusion about roompc.

Table 4 and Fig. 1 report the estimates for the three indicators for the eight considered years for both rural and urban areas. Before to analyze the results, it is important to note as the estimates for 2000 and 2007 (in italics) are design-based estimates and that inside the brackets, under each estimate, we report the standard deviation. According with all the indicators the inequality is higher in the urban areas than in the rural ones.

### 5.1. Inequality and climate changes: is there a relation?

Table 4 reports estimates of three different inequality indicators which progress from 2000 to 2007 seems to be not linear. In fact, it is possible to appreciate as every inequality indicator for both rural and urban areas decrease from 2001 to 2003 and then returned to growth until 2007 where levels of inequality equal or greater than those of 2000 were reached. To understand this trend, it is necessary to look at the geography and at the infrastructure of Morocco.

Simulations regarding future conditions of North-Africa indicate they will experience a rapid increase in temperature and an even more erratic rainfall pattern [20]. As noted by Alfani et al. [21] poor households secure their livelihood mainly through rain-fed agriculture and on-farm activities. Where these become irregular or no longer sufficient for agriculture obviously inequality within the state will increase. It becomes more and more important to quantify the risks that endanger nutrition for these individuals/households and tailor policy interventions, which include prevention and mitigation strategies.

As the Moroccan economy highly depends on agriculture, we expect a significant correlation between rainfall and inequality. In periods of drought, we expect inequality to increase since most of the agriculture is rainfed and only a few very developed areas (for example the Settat area) are equipped with modern irrigation systems that enable farmers to withstand shocks, which exacerbate rural inequality. In periods of abundant rainfall, we anticipate better performance overall of the agricultural economy and thus a decline in inequality.

**Table 4**
Estimates of Gini, Theil and Atkinson index for Moroccan rural/urban areas from 2000 to 2007. Standard deviations of the estimates are reported in brackets.

| Years | Gini Index | | Theil Index | | Atkinson Index | |
|---|---|---|---|---|---|---|
| | Rural | Urban | Rural | Urban | Rural | Urban |
| *2000* | *0.349* | *0.408* | *0.240* | *0.325* | *0.101* | *0.136* |
| | *(0.037)* | *(0.051)* | *(0.017)* | *(0.016)* | *(0.012)* | *(0.018)* |
| 2001 | 0.321 | 0.414 | 0.267 | 0.347 | 0.116 | 0.146 |
| | (0.037) | (0.041) | (0.027) | (0.035) | (0.119) | (0.149) |
| 2002 | 0.303 | 0.356 | 0.184 | 0.308 | 0.089 | 0.101 |
| | (0.033) | (0.358) | (0.018) | (0.021) | (0.009) | (0.010) |
| 2003 | 0.308 | 0.422 | 0.154 | 0.318 | 0.109 | 0.152 |
| | (0.036) | (0.042) | (0.027) | (0.381) | (0.011) | (0.015) |
| 2004 | 0.329 | 0.427 | 0.182 | 0.376 | 0.131 | 0.154 |
| | (0.039) | (0.043) | (0.037) | (0.037) | (0.013) | (0.015) |
| 2005 | 0.336 | 0.425 | 0.191 | 0.345 | 0.143 | 0.152 |
| | (0.041) | (0.048) | (0.037) | (0.027) | (0.064) | (0.014) |
| 2006 | 0.328 | 0.423 | 0.213 | 0.359 | 0.140 | 0.151 |
| | (0.071) | (0.019) | (0.016) | (0.026) | (0.012) | (0.021) |
| *2007* | *0.349* | *0.426* | *0.225* | *0.357* | *0.102* | *0.151* |
| | *(0.077)* | *(0.049)* | *(0.035)* | *(0.022)* | *(0.010)* | *(0.021)* |

To verify this hypothesis, i.e. that the decline in the inequalities between 2001 and 2003 is due to more favorable weather condition, we compare inequality measures areas with average annual precipitation data published by the World-Bank climate knowledge portal, for both rural and urban areas. The average annual precipitations are measured in millimeters per year. In Fig. 1, alongside the estimates presented in Table 4, the average annual precipitations are graphed. The inverse correlation between indicators of inequality and precipitation becomes apparent at first glance.

First of all, we use the correlation coefficient testing if there is a negative correlation between the inequality measures and the average annual precipitation. Table 5 reports the results showing as there is a negative correlation statistically different from zero for all the three indicators for the rural areas. It is not surprising that the correlations are not statistically lower than 0 in the urban areas for the Gini and the Atkinson index where agricultural is less important. To test the causalities between the indicators and the precipitation we use the Granger causality test. As results we obtain that there is a real causality and not a spurious correlation between precipitation and inequality indicators. To conclude, we highlight as similar results are obtained for the same years by Bijaber et al. [22] who note as the import of cereals in Morocco is positively related to drought.

In Morocco, where structural transformation of the economy is slow, agriculture still employs a relevant part of the population, often the poorest. Unfavorable climatic shocks such as drought tend to have an asymmetric impact and affect the more vulnerable parts of the population, this clearly, all other things being equal, tend to increase in inequality. Our study offers new evidence to help policy makers to assess the impact of climatic shocks and adequately address them. For example, by creating safety nets that protect vulnerable farmers from droughts it can mitigate their negative impact and consequently contribute to minimize the increase of inequality and poverty that these can cause.

## 6. Simulation set-up

Simulations are frequently used to determine estimators' characteristics when it is hard to achieve analytic results on estimators' properties. To do it we generate a "two-type" simulation. As known simulations are roughly divided into two sub-groups model- and design-based [23]. In SSIT context it is impossible to implement a complete design-based simulation given the impossibility to have the dependent variable for years between $t_-$ and $t_+$. To perform both a design- and a model-based simulations we propose the following steps:

1 According with HBS surveys, for each available year, chose the probabilistic distribution $\mathscr{F}$ that best fit our data;
2 Estimated the model (2) for time $t_-$ and $t_+$ using as covariates sex, age and random effects based on urban-rural areas obtaining $\mathscr{F}(\mu_{ij}^{t_-}, \sigma_{ij}^{t_-}, \nu_{ij}^{t_-}, \tau_{ij}^{t_-})$ and $\mathscr{F}(\mu_{ij}^{t_+}, \sigma_{ij}^{t_+}, \nu_{ij}^{t_+}, \tau_{ij}^{t_+})$;
3 Create a model-based population for each year between $t_-$ and $t_+$ using model (3) and covariates taken from LFS surveys;
4 For each year we repeatedly select (500 times) samples by stratified sampling with region acting as strata. The sample sized is equal to the 3 % of the original size.

In this way we basically have a design-based simulation for years $t_-$ and $t_+$ and a model-based simulation for each year between $t_-$ and $t_+$.

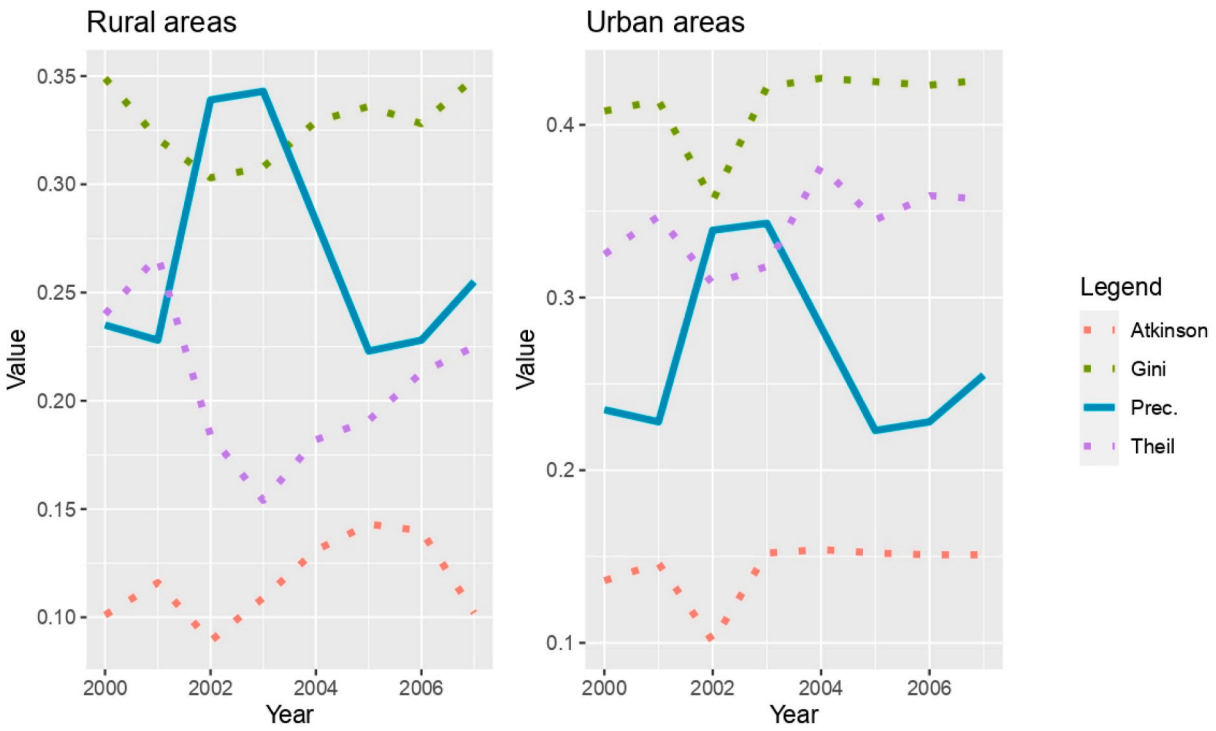


**Fig. 1.** Estimates of Gini, Theil and Atkinson index for Moroccan rural/urban areas from 2000 to 2007 and mean precipitations. Mean precipitations are reported in mm per year/1000.

**Table 5**
Estimated coefficient of correlation between Gini, Theil and Atkinson index and average annual precipitation for Moroccan rural/urban areas. P-value in brackets.

| | Rural | Urban |
|---|---|---|
| Gini Index | −0.727 (0.020) | −0.499 (0.103) |
| Theil Index | −0.730 (0.019) | −0.542 (0.082) |
| Atkinson Index | −0.537 (0.085) | −0.459 (0.126) |

We study characteristics and performance of SSIT-GAMLSS and SSIT for the estimation of three different indicators: Theil index, Gini index and Atkinson index (aversion parameter equal to 0.5). In particular we have $t_- = 2000$ and $t_+ = 2007$. Results are evaluated both in terms of bias and variability. To measure the bias, we consider the Relative Bias (RB) and to measure the variability both the MSE and the Coefficient of Variation (CV). At the step 2 of the simulation set-up and according with Table 1 we use the GB2 distribution.

### 6.1. Simulation results

As expected, differences between SSIT-GAMLSS and SSIT are significant. Let us start from the $R^2_{adj}$. The mean $R^2_{adj}$ for the SSIT is equal to 0.08 for 2000 and 0.11 for 2007 while for the SSIT-GAMLSS the mean $R^2_{adj}$ increase to 0.19 and 0.21, respectively. Those differences in the coefficient of determination are the results of the choice of the GB2 distribution instead of the normal distribution of the SSITs. Fig. 2 reports the estimates for the urban-rural area for each year and for each indicator. While, Table 6 reports the mean of the RB, MSE and CV for each index for every simulation. What it is important to note it is that SSIT-GAMLSS reduce the RB in every single scenario when compared with SSIT. SSIT, as known and as explained in the previous section, tends to severally underestimate inequality indicators. The RB reduction is mostly given by the properties of the GAMLSS. GAMLSS, in-fact, belong to the so called beyond mean regression models [24]. These types of regressions using different distribution from the normal one allowing for the use of covariates in every parameter of the distribution are able to predict values far from the mean. One of the most known problems of the linear regression and so of the SSIT is the so call regression to the mean. Results from the simulation clearly (Fig. 2) highlight this problem. Computing the variance of the estimates obtained with SSIT for each indicator we obtain results on the order of $10^{-9}$ which, basically, means that with SSIT we find at every year for each indicator the same value, the mean one. The same value computed for SSIT-GAMLSS reduce to only $10^{-3}$ which, again, highlight the ability of GAMLSS to predict specific value year by year far from the mean one. Moving to the MSE it is possible to note the SSIT-GAMLSS is also more accurate than the SSIT. This is also appreciable form the CVs. It is important to note that the highest value of both the MSE and the CV for 2000 and 2007 are to be imputed to two main reasons: 2000 and 2007 are to be considered as design-based simulations which usually have higher MSE/CV than the model one and those years are also the one with less sampled units. In fact, as said in section 4 the numerosity of the HBSs is lower than the one of LFSs.

## 7. Concluding remarks

This paper aims to contribute to the ongoing debate on ways to improve the accuracy and timeliness of welfare statistics in developing countries experiencing severe budget constraints. Only a few developing countries have the capacity to collect annual data on income or expenditure, therefore, indicators such as inequality or poverty rates can be computed in these countries only when Household Budget Surveys are available, about every four to five–and sometimes as many as seven–years. To overcome this problem, methods have been developed to compare these indicators over time from surveys that are little comparable. These techniques (SSITs) have proven effective at predicting poverty indicators but are much less accurate when used for inequality indicators.

To illustrate this limitation, we conducted a simulation based on data from Moroccan Household Budget Surveys and Labor Force Surveys.

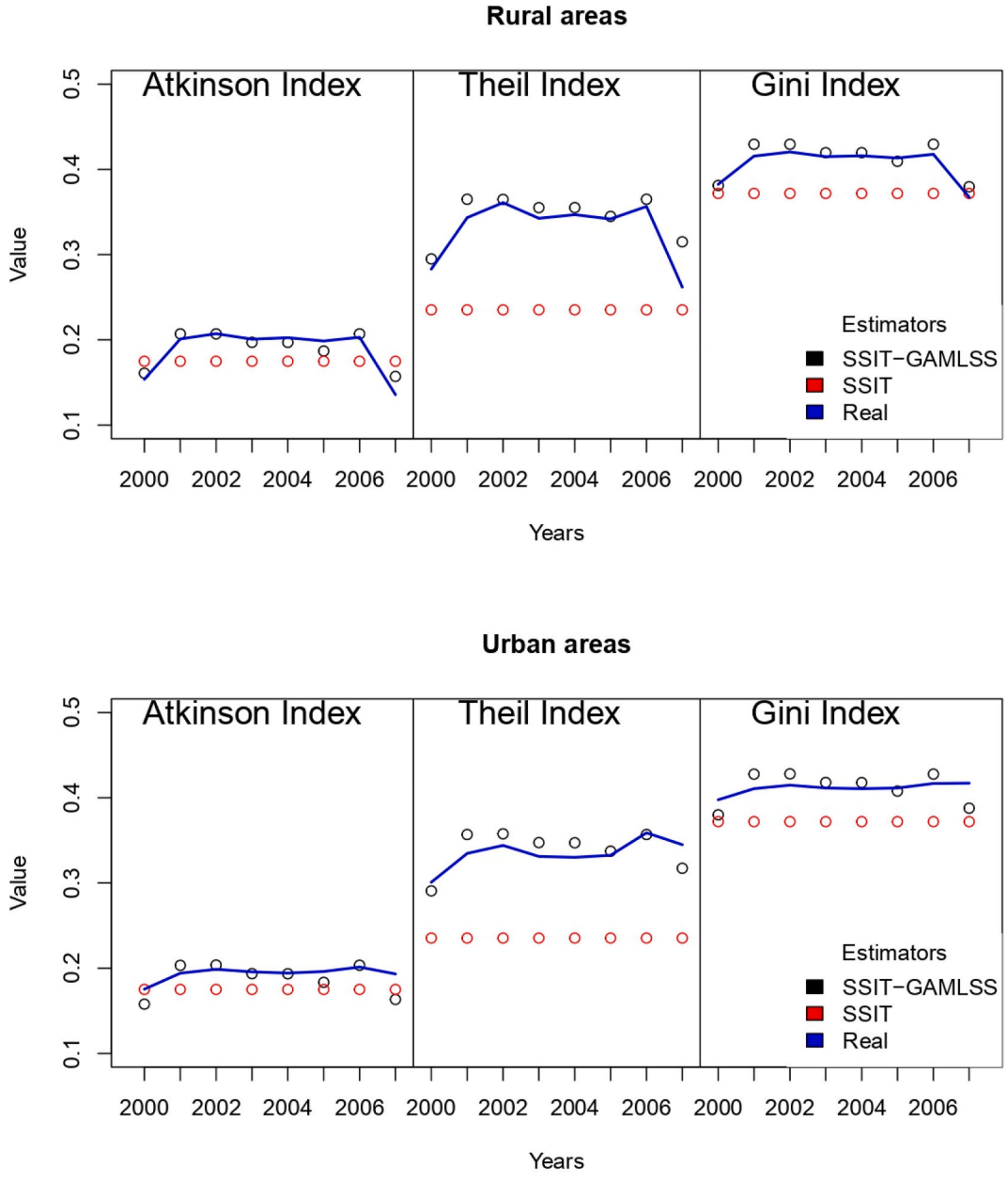


**Fig. 2.** Simulation results: Estimates with SSIT-GAMLLS and SSIT compared with real values for rural and urban areas.

**Table 6**
Simulation results: relative bias, mean square error and coefficient of variation for SSIT-GAMLSS and SSIT.

| | SSIT-GAMLSS | | | SSIT | | |
|---|---|---|---|---|---|---|
| | Design-based | Model-based | Averall | Design-based | Model-based | Averall |
| Atkinson Index | | | | | | |
| Average RB | 0.042 | −0.004 | 0.007 | 0.088 | −0.067 | −0.028 |
| Average MSE | 0.004 | 0.0001 | 0.001 | 0.005 | 0.0005 | 0.001 |
| Average CV | 0.059 | 0.003 | 0.017 | 0.192 | 0.019 | 0.062 |
| Theil Index | | | | | | |
| Average RB | 0.122 | 0.027 | 0.051 | −0.135 | −0.325 | −0.277 |
| Average MSE | 0.0002 | 0.0001 | 0.0007 | 0.006 | 0.0002 | 0.002 |
| Average CV | 0.055 | 0.003 | 0.016 | 0.329 | 0.069 | 0.134 |
| Gini Index | | | | | | |
| Average RB | 0.014 | 0.015 | 0.015 | −0.010 | −0.107 | −0.082 |
| Average MSE | 0.186 | 0.0007 | 0.046 | 0.192 | 0.003 | 0.048 |
| Average CV | 0.819 | 0.002 | 0.206 | 0.882 | 0.016 | 0.232 |

Our results indicate that the predicted inequality measures are severally negative biased. Our theoretical explanation for this points to two main limitations of the Standard SSITs models based on linear regressions: the overly stringent assumption of residuals normality distribution and the expectation that regression-based models predict distribution compressed around the mean and with thin tails. Unfortunately, the shape of the tails is crucial to correctly estimate inequality. Thus, almost by design, these models tend to produce estimates that are far below the correct values.

The method we propose is based on generalized additive models for location, scale and shape which not only release the normality assumption but also allow researcher to explain with covariates every parameter of a distribution and not only the location one. With this algorithm, we reduce the bias and obtain results that are not systematically biased. Furthermore, the estimates of inequality indices for the years in which only labor force data are available seem to be consistent with Moroccan economic trends. Moreover, we prove the relation between the inequalities' trends from 2000 to 2007 with the precipitations. Although those may seem like insignificant results it is important to highlight as without SSIT-GAMLSS this relation will be impossible to be verified and, as the knowledge of the causalities can help policymakers to prevent increase in the inequalities such as implementing with modern irrigation systems more rural areas. Moreover, a further development could be the use of unit-level variables (as done in this paper) together with area-level variable. In other words, between the regression coefficients can be added also covariates at the same level of the random effects. Those covariates could take the same value for every unit in the same area and can be taken from meteorological data such as the rainfall variable.

## CRediT authorship contribution statement

**Gianni Betti:** Writing – original draft, Writing – review & editing. **Vasco Molini:** Data curation, Writing – original draft, Writing – review & editing. **Lorenzo Mori:** Methodology, Software, Writing – original draft, Writing – review & editing.

## Data availability

The data that has been used is confidential.

## References


[1] Elbers C, Lanjouw JO, Lanjouw P. Micro-level estimation of poverty and inequality. Econometrica 2003;71(1):355–64.
[2] Mathiassen A. Testing prediction performance of poverty models: Empirical evidence from Uganda. Rev Income Wealth 2013;59(1):91–112.
[3] Elbers C, Lanjouw PF, Mistiaen JA, Özler B, Simler K. On the unequal inequality of poor communities. World Bank Econ Rev 2004;18(3):401–21.
[4] Dabalen A, Graham EG, Himelein K, Mungai R. Estimating poverty in the absence of consumption data: the case of Liberia. World Bank Policy Research Working Paper; 2014. 7024.
[5] Dang H-A, Jolliffe D, Carletto C. Data gaps, data incomparability and data imputation: a review of poverty measurement methods for data-scarce environments. J Econ Surv 2019;33(3):757–97.
[6] Demombynes G, Hoogeveen JG. Growth, inequality and simulated poverty paths for Tanzania, 1992–2002. J Afr Econ 2007;16(4):596–628.
[7] Newhouse DL, Shivakumaran S, Takamatsu S, Yoshida N. How survey-to- survey imputation can fail. World Bank Policy Research Working Paper; 2014. p. 6961.
[8] Douidich M, Ezzrari A, Van der Weide R, Verme P. Estimating quarterly poverty Rates using labor force surveys: a primer. World Bank Econ Rev 2016;30(3):475–500.
[9] Schluter C. On the problem of inference for inequality measures for heavy-tailed distributions [Publisher: Oxford University Press, Oxford, UK] Econom J 2012;15(1):125–53.
[10] More S. Identifying and overcoming transformation bias in forecasting models. ArXiv preprint arXiv:2208.12264; 2022.
[11] Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. J Roy Stat Soc: Series C (Applied Statistics) 2005;54(3):507–54.
[12] Graf M, Nedyalkova D. Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. Rev Income Wealth 2014;60(4):821–42.
[13] Chotikapanich D, Griffiths WE, Hajargasht G, Karunarathne W, Rao DP. Using the GB2 income distribution. Econometrics 2018;6(2):21.
[14] Newhouse DL, Vyas P. Nowcasting poverty in India for 2014-15: a survey to survey imputation approach. World Bank Policy Research Working Paper; 2018.
[15] Ketkar KW, Ketkar SL. Socio-demographic dynamics and household demand [Publisher: JSTOR]. E Econ J 1987;13(1):55–62.
[16] Stasinopoulos DM, Rigby R, Heller G, Voudouris V, De Bastiani F. Flexible regression and smoothing: using GAMLSS in R. Chapman & Hall CRC; 2017.
[17] Rigby RA, Stasinopoulos DM, Heller GZ, De Bastiani F. Distributions for modeling location, scale, and shape: using GAMLSS. first ed. Chapman; Hall; 2019.
[18] Rust Keith F, Rao JNK. Variance estimation for complex surveys using replication techniques. Stat Methods Med Res 1996;5(3):283–310.
[19] Field CA, Welsh AH. Bootstrapping clustered data. J Roy Stat Soc B 2007;69(3):369–90.
[20] Field CB, Barros VR. Climate change 2014–impacts, adaptation and vulner- ability: regional aspects. Cambridge University Press; 2014.
[21] Alfani F, Dabalen A, Fisker P, Molini V. Vulnerability to stunting in the west african sahel. Food Pol 2019;83:39–47.
[22] Bijaber N, El Hadani D, Saidi M, Svoboda MD, Wardlow BD, Hain CR, Poulsen CC, Yessef M, Rochdi A. Developing a remotely sensed drought monitoring indicator for Morocco. Geosciences, MDPI 2018;8(2).
[23] Temple M. Simulation for data science with R. Packt Publishing Ltd; 2016.
[24] Kneib T. Beyond mean regression. Stat Model Int J 2013;13(4):275–303.
[25] Clementi F, Fabiani M, Molini V, Schettino F, Kahn H. Polarization and its discontents: Morocco before and after the arab spring». J Econ Inequal 2022;21(1):105–29.
[26] Krafft C, Assaad R, Nazier H, Ramadan R, Vahidmanesh A, Zouari S. Estimating poverty and inequality in the absence of consumption data: an application to the middle east and north Africa. Middle East Dev J 2019;11(1):1–29.



**Gianni Betti** is Full Professor in statistics and economics at the Department of Economics and Statistics, University of Siena (Italy). He has worked on several projects for the World Bank and European Commission and has been closely involved with the department of the EU Statistics on Income and Living Conditions.

**Vasco Molini** is Program manager at the World Bank. Over the course of more than one decade with the World Bank, he worked in Mozambique, Angola, Sao Tome and Principe, Ghana, Nigeria, Morocco, Tunisia and Libya, where he led country level dialogues on jobs, poverty reduction, polarization and conflicts; and managed country-level statistical capacity building initiatives. Vasco has also published on these topics in more than twenty internationally peer-reviewed journals. He holds a PhD in Economics from the University of Florence and a Post-Doc from the Free University of Amsterdam.

**Lorenzo Mori** is a PhD student is Statistics at the University of Bologna, Italy. He collaborates with Professor Gianni Betti of the University of Siena, Italy, in his studies on poverty and living conditions. He is interested in economic and social statistics.