



Composite-Based Path Modeling for Conditional Quantiles Prediction. An Application to Assess Health Differences at Local Level in a Well-Being Perspective

Cristina Davino¹ · Pasquale Dolce² · Stefania Taralli³ · Domenico Vistocco⁴

Accepted: 28 June 2020 / Published online: 24 July 2020
© The Author(s) 2020

Abstract

Quantile composite-based path modeling is a recent extension to the conventional partial least squares path modeling. It estimates the effects that predictors exert on the whole conditional distributions of the outcomes involved in path models and provides a comprehensive view on the structure of the relationships among the variables. This method can also be used in a predictive way as it estimates model parameters for each quantile of interest and provides conditional quantile predictions for the manifest variables of the outcome blocks. Quantile composite-based path modeling is shown in action on real data concerning well-being indicators. *Health* outcomes are assessed taking into account the effects of *Economic well-being* and *Education*. In fact, to support an accurate evaluation of the regional performances, the conditions within the outcomes arise should be properly considered. Assessing health inequalities in this multidimensional perspective can highlight the unobserved heterogeneity and contribute to advances in knowledge about the dynamics producing the well-being outcomes at local level.

Keywords PLS path modeling · Quantile composite-based path modeling · Conditional quantile prediction · Well-being · Territorial inequalities · Health indicators

✉ Cristina Davino
cristina.davino@unina.it

Pasquale Dolce
pasquale.dolce@unina.it

Stefania Taralli
taralli@istat.it

Domenico Vistocco
domenico.vistocco@unina.it

¹ Department of Economics and Statistics, University of Naples Federico II, Naples, Italy

² Department of Public Health, University of Naples Federico II, Naples, Italy

³ ISTAT, Roma, Italy

⁴ Department of Political Science, University of Naples Federico II, Naples, Italy

1 Introduction

This paper outlines the methodology and the results of a study concerning the relationships among three well-being domains (*Education, Economic Well-Being and Health*) measured on Italian provinces. Data come from the Italian system of indicators on *Equitable and Sustainable Well-Being (Benessere Equo e Sostenibile—BES)* proposed by the National Institute of Statistics (ISTAT 2018). BES represents a well established reference database in the national and international debate on the research of alternative well-being measures. The present paper proposes an advancement of the work elaborated in Davino et al. (2017, 2018), where a hierarchical composite model was used to study the relationships among components of the BES. The proposal exploits quantile regression (QR) (Koenker and Basset 1978; Koenker 2005) to obtain the best predictions in a network of simultaneous equations.

The introduction of structural equation modeling has been a turning point in the analysis of complex relationships among unobservable variables, first in its hard modeling, or covariance-based approach (Jöreskog 1978), and then in the soft modeling, or composite-based approach (Wold 1982; Tenenhaus et al. 2005). This paper embraces the soft modeling approach, which does not require any distributional assumption on the variables and exploits non-parametric methods to estimate the model parameters. Partial Least Squares Path Modeling (*PLS-PM*), proposed by Wold (1985), Tenenhaus et al. (2005), is one of the most widespread composite-based method for structural equation modeling. Nowadays, *PLS-PM* is a well established method both in statistical literature (Esposito Vinzi et al. 2010) and in applied research in several disciplines (Henseler et al. 2009; Hair et al. 2012; Sarstedt et al. 2017; Di Napoli et al. 2019).

Recently a quantile approach to *PLS-PM* called Quantile Composite-based Path Modeling (*QC-PM*) was proposed by Davino and Esposito Vinzi (2016) to broaden the potential of *PLS-PM*. To this end, *QC-PM* exploits QR (Koenker and Basset 1978; Koenker 2005) in all the steps of the *PLS-PM* estimation algorithm. This allows to highlight if and how the relationships among variables change according to the explored quantile of interest. It is worth to emphasize that *QC-PM* is not an alternative to *PLS-PM*, but rather its ideal completion. *PLS-PM* aims to estimate the effect of the involved variables on the conditional mean of the responses, *QC-PM* extends the focus to the whole conditional distribution; *PLS-PM* provides an effective summary of the dependence structure, *QC-PM* is a useful tool to magnify it.

The present study deals with the use of composite-based models for predictive purposes. In fact, *PLS-PM* cannot easily be used as a predictive modeling because the network of relationships is complex and the identification of a single direction to be explained is troublesome. At this regard, see Evermann and Tate (2016), Shmueli et al. (2016) and Dolce et al. (2017). Starting from the explicit and general formulation of the predictive model introduced by Dolce and Hanafi (2017) and used in Dolce et al. (2018), the present paper proposes a predictive-oriented *QC-PM*, namely a model able to provide the best prediction for each statistical unit. It extends the approach proposed by Davino and Vistocco (2015, 2018) which was aimed at identifying a typology in a dependence model. The authors introduced the “best quantile” for each unit, i.e. the quantile associated with the conditional model that provides the best estimate of the response variable (the best model for each unit). This paper exploits the “best quantile” approach in the composite-based path modeling in order to obtain accurate predictions by estimating several path models at different quantiles.

Even if the risk of overfitting is always lurking when we deal with prediction, it does not actually matter for the case concerned here since we aim to provide the best in-sample predictions and not to generalize on different data. The analyzed dataset contains indeed all the population units, namely the Italian provinces, and we aim to define the most accurate model for each *Health* indicator and the best predictive model for each province.

2 The Reference Framework

The empirical application focuses on the prediction of health levels and health inequalities in a regional well-being perspective. BES is the reference framework. It consists of a wide set of about 130 statistical indicators produced by the Italian National Institute of Statistics (Istat) to describe and monitor the progress of Italian society from a social and environmental point of view in a comprehensive way (ISTAT 2013, 2018).

Health is one of the 12 domains of well-being considered in the BES framework, together with *Education and training*, *Work and life balance*, *Economic well-being*, *Social relationships*, *Politics and Institutions*, *Safety*, *Subjective well-being*, *Landscape and cultural heritage*, *Environment*, *Innovation, research and creativity*, and *Quality of services*.

In the BES framework, *Health* is seen as a central element in life and an essential condition for people's well-being and prosperity of populations. In fact, *Health* outcomes are related to many dimensions of the individual and social well-being. Among the multiple relationships, which link *Health* to the other BES domains and assets, we focus here on two related domains: *Education and training* and *Economic well-being*. In a well-being perspective, *Education* does not only have an intrinsic value but it directly affects other well-being domains. People with higher education levels have higher standards of living and more possibilities to find work, they live longer and better because they have healthier lifestyles, easier access to services and more opportunities to find less risky jobs. Similarly, *Economic well-being* is both an asset of BES and a driver of the well-being outcomes in other domains. Indeed, earning capacities and economic resources ensure that an individual can obtain and support a specific standard of living.

Recent studies (Costa et al. 2014; ISTAT 2019b; Murtin et al. 2017; Petrelli et al. 2019) focused on the relationships between health and socio-economic conditions at individual level. They confirm that regional disparities in health outcomes are still marked in Italy, both in terms of life expectancy at birth and mortality risk. Health inequalities among Italian regions arise regardless of age, gender and socioeconomic status, but they clearly appear to be related to socio-economic factors, as they have a higher impact in the poorer southern regions of Italy. Furthermore, lower education levels explain a considerable proportion of mortality risk, although with different effects by geographical area: males with a lower education level throughout Italy have a life expectancy at birth that is 3 years less than those with higher education; residents in southern Italy lose an additional year in life expectancy, regardless of education level. Other studies highlight that health inequalities are more severe within the southern Italian regions than within the northern ones. The single and joint effects of education and income factors are remarkable: mortality inequalities between better educated and less educated people explain globally about 25% of deaths among men and more than 10% among women. The differences in health outcomes among the Italian regions also result from local policies, as the Italian Regional Administrations have the main power to regulate and organize the public health services; instead, focusing on the sub-regional level, the other differentiating factors gain greater importance.

Therefore, the impact of economic factors and of education levels is bearer of interesting analysis. In particular, it would be useful to obtain a model to predict health outcomes controlling for the factors that affect *Education* and *Economic Well-Being*.

3 Methodological Framework

3.1 Partial Least Squares Path Modeling

PLS-PM, originally developed by Wold (1982, 1985), is a powerful multivariate statistical method that can be applied to the study of the relationships among K blocks of observed variables. Such blocks of observed or manifest variables (MVs), $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_K]$, measure K latent variables (LVs), ξ_1, \dots, ξ_k , usually named as components or composites. *PLS-PM* is commonly considered an alternative approach to the covariance structure analysis (Jöreskog 1978) of Structural Equation Modeling (*SEM*), although these two approaches belong to two different families of statistical methods.

PLS-PM focuses on LV scores computation, accounting for variances of MVs and correlations between LVs. Each block of MVs is summarized in a component, or a composite (i.e. an exact linear combination of the MVs), that maximizes the explained variance of the set of MVs. Therefore, *PLS-PM* is commonly referred to as a component-based, composite-based or variance-based approach. Great flexibility, robustness, few demands concerning distributional assumptions and requirement for identification are the main features of *PLS-PM*, and underpin its widespread dissemination in many areas (Esposito Vinzi et al. 2010; Hair et al. 2014).

More formally, let us consider that P variables are collected in a table \mathbf{X} of data partitioned in K blocks:

$$\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_J, \mathbf{X}_{J+1}, \dots, \mathbf{X}_{J+Q}, \mathbf{X}_{J+Q+1}, \dots, \mathbf{X}_K]$$

Let $\mathbf{X}_k = \{x_{ip_k}\}$ be the generic block, where

- the input blocks are in the first J positions,
- the intermediate blocks run from block $J + 1$ to block $J + Q$,
- the output blocks run from block $J + Q + 1$ to block K
- $i = 1, \dots, n$, with n denoting the number of observations,
- $p_k = 1, \dots, P_k$, with P_k being the number of MVs in the k -th block.

We denote by $\xi_k = \{\xi_{ik}\}$ the corresponding LVs for each block of variables. A generic MV is instead denoted by $\mathbf{x}_{p_k} = \{x_{ip_k}\}$. The path diagram in Fig. 1 shows an example of a simple path model with an input, an intermediate and an output block of manifest variables.

The general model consists of two sub-models: the inner (or structural) model and the outer (or measurement) model. The measurement model relates each MV to its own LV by the following equation:

$$\mathbf{x}_{p_k} = \lambda_{pk0} + \lambda_{pk} \xi_k + \epsilon_{pk}, \tag{1}$$

where λ_{pk0} is a location parameter, λ_{pk} is the loading coefficient that captures the effect of ξ_k on \mathbf{x}_{p_k} , and ξ_k is the measurement error variable. The structural model captures and specifies the dependence relationships among LVs. A generic dependent LV is linked to the corresponding explanatory LVs by the following model:

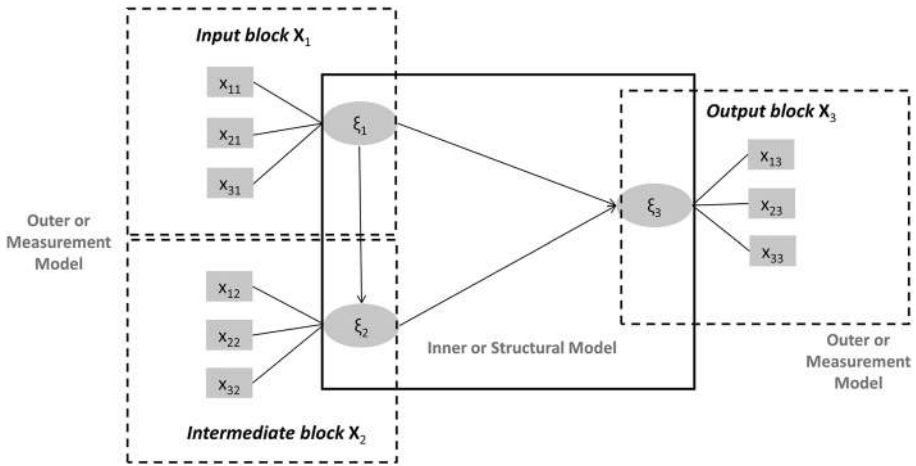


Fig. 1 A path model with an input, an intermediate and an output block of manifest variables

$$\xi_k = \beta_{k0} + \sum_{k' \rightarrow k} \beta_{kk'} \xi_{k'} + \zeta_k, \tag{2}$$

where $\beta_{kk'}$ is the so-called path coefficient that captures the effects of the predictor LV $\xi_{k'}$ on the dependent LV ξ_k , and ζ_k is the inner residual variable.

The following weighting relation defines the casewise scores of each LV as a weighted aggregate of its own MVs:

$$\hat{\xi}_k = \sum_{p=1}^{P_k} \hat{w}_{kp} \mathbf{x}_{pk}, \tag{3}$$

where \hat{w}_{kp} is the outer weight obtained through the *PLS-PM* iterative algorithm.

Since in *PLS-PM* there are different kinds of residual variables, a set of partial (or local) least squares (PLS) criteria are defined and the optimal solution is found by an iterative algorithm (Lohmöller 1989). In particular, the estimation of the model parameters in Equations (1) and (2) proceeds in two stages. The first stage computes the outer weight vectors \hat{w}_k in Equations (3), and consequently the composite ξ_k , through an iterative algorithm alternating OLS simple or multiple regressions. The second stage estimates the loading coefficients λ_{pk} and the path coefficients $\beta_{kk'}$ through classical OLS regressions.

The statistical and numerical properties of *PLS-PM* were deeply investigated and interesting results were found in terms of global optimization criteria and convergence properties (Glang 1988; Mathes 1993; Hanafi 2007; Krämer 2007; Tenenhaus and Tenenhaus 2011). Furthermore, recent methodological developments introduced interesting features of the method, starting on which it is possible to generate predictions from PLS path models. The following subsection details prediction through *PLS-PM*.

3.2 Predictive-Oriented PLS-PM

PLS-PM is a powerful method both for explorative and predictive purposes. This is a distinctive feature compared to covariance-based SEM, which mainly focuses on obtaining

valid inferences for population parameters. However, since its origin *PLS-PM* has been almost exclusively used as an explanation oriented technique. Only in the recent years the predictive ability of *PLS-PM* started to gain increasing interest from researchers (Evermann and Tate 2016; Shmueli et al. 2016; Dolce et al. 2017; Danks and Ray 2018; Shmueli et al. 2019; Sharma et al. 2019). The limitation of the *PLS-PM* as explanatory modeling was due to the lack of an explicit formulation of the predictive model, because of the complexity of the PLS path model. Two models are considered in *PLS-PM*, i.e. Equations (1) and (2), and data are partitioned into three kinds of blocks: input (only used for prediction), intermediate (used for prediction and as dependent blocks) and output blocks (only used as dependent blocks). As a matter of fact, prediction in *PLS-PM* has been considered a difficult task because a choice should be made between either prediction from the measurement model or from the structural model. Moreover, prediction of individual observations may refer to either individual LV score observations or individual observations of the MVs in the dependent blocks. Finally, intermediate blocks pose a special challenge in the predictive context, because they play a twofold role in the model: they are both predictor variable blocks and dependent variable blocks.

Lohmöller (1989) defines five different sorts of predictions from *PLS-PM*:

1. *communality prediction*: each MV is predicted by the corresponding LV—Equation (1);
2. *structural prediction*: the prediction of each LV is obtained using the related predictor LVs—Equation (2);
3. *validity prediction*: the prediction of each LV is obtained using their MVs—Equation (3);
4. *redundancy prediction*: each MV is predicted by the predictor LVs that is directly connected to its own LV;
5. *operative prediction*: each MV is predicted using only the MVs of the predictor blocks (all the LVs are replaced with their corresponding weight relation)—Equation (3).

Despite this complexity, it is possible to generate predictions from *PLS-PM* since appropriate schemes were recently proposed Shmueli et al. (2016), Dolce and Hanafi (2017). The present paper uses the explicit and general formulation of the predictive model proposed in Dolce and Hanafi (2017), which incorporates both the measurement and structural model in an unique model, and requires only MVs as predictors and outcomes, following the operative prediction defined in Lohmöller (1989).

3.3 Quantile Composite-Based Path Modeling

QC-PM has been proposed by Davino and Esposito Vinzi (2016) to complement *PLS-PM*. *QC-PM* exploits QR to explore the whole distribution of dependent variables as function of the set of predictors. Since *PLS-PM* is based on simple and multiple OLS regressions, its coefficients focus on the conditional means of the dependent variables. Even if it provides an effective summary, in some cases the estimates of coefficients may vary along the distribution of the dependent variable. This happens in presence of heteroscedastic variances of the errors or highly skewed dependent variables. In such cases, *PLS-PM* may give an incomplete picture of the relationships among variables. The quantile approach is instead able to model the location, the scale and the shape of the responses.

QR was introduced by Koenker (2005) to extend the regression model from the conditional mean to any conditional quantile of interest. In linear models, QR estimates have

the same interpretation of any other linear model. The intercept measures the response value when all the regressors are set to zero, the slopes measure the rates of change in the response per a unit change in the value of the correspondent regressor, keeping all the others constant. Since QR estimates a set of coefficients (intercept and slopes) for each considered quantile, coefficients must be interpreted in terms of the quantiles of the response. A dense grid of equally spaced quantiles can provide a fairly accurate approximation of the whole quantile process (Furno and Vistocco 2018) and a reconstruction of the whole conditional distribution of the response variable, because each conditional quantile predicts the correspondent location of the response variable (Davino et al. 2013).

QC-PM uses QR instead of OLS regression in all the estimation steps. In particular, QC-PM consists in introducing a QR in the steps described in Sect. 3.1 to estimate model parameters—Equations (1), (2) and (3). For each quantile θ of interest, $\theta \in (0, 1)$, the first stage computes the outer weights vectors w_k and each composite ξ_k , through an iterative algorithm alternating simple QR or multiple QR. The second stage estimates the loading coefficients $\lambda_{pk}(\theta)$ and the path coefficients $\beta_{kk'}(\theta)$ through QR:

$$x_{pk}(\theta) = \lambda_{pk0}(\theta) + \lambda_{pk}(\theta)\xi_k + \epsilon_{pk}, \tag{4}$$

$$\xi_k(\theta) = \beta_{k0}(\theta) + \sum_{k' \rightarrow k} \beta_{kk'}(\theta)\xi_{k'} + \zeta_k. \tag{5}$$

Similarly to the unconditional quantile minimisation (Fox and Rubin 1964), the conditional quantile estimator is obtained by minimizing a weighted sum of residuals. The function to minimize in case of equation (4) is:

$$\hat{\lambda}(\theta) = \underset{\lambda(\theta)}{\operatorname{argmin}} \sum_{i=1}^n \rho_{\theta}(x_i - \xi_i^T \lambda(\theta)) \tag{6}$$

where the first element of ξ is equal to 1 to include the intercept, and $\rho_{\theta}(\cdot)$ is the check function, which asymmetrically weights positive and negative residuals, namely:

$$\rho_{\theta}(u) = \begin{cases} \theta u & \text{if } u > 0 \\ (1 - \theta)u & \text{if } u \leq 0 \end{cases} \tag{7}$$

The same holds in case of Equation 5.

For each quantile of interest, QC-PM provides a set of outer weights, loadings and path coefficients. Therefore, it offers a more complete picture of the relationships among variables both in the outer model (as the outer weights measure the effects of each MV on the corresponding construct) and in the inner model (as the path coefficients quantify the impact of lower-order constructs on higher order constructs). However, in order to compare path coefficients estimated over quantiles, measurement invariance has to be satisfied in the models (Henseler et al. 2016). In other words, for each MV, all the loading should be very similar across quantiles and compared to the one estimated by PLS-PM, because the same LV should be measured across quantiles. If loadings change across quantiles, there is no guarantee that a LV is measuring the same concept and path coefficients estimated at different quantiles cannot be reliably compared.

In this situation, a possible solution may be to fix the quantile to the median in the measurement model and letting quantiles change just in the structural part, following the approach proposed by Wang et al. (2016) in factor-based structural equation modeling. This approach is justified if we consider that the role of the measurement model is to relate

the MVs to LV and to construct a score for the LV. Finally, measures of goodness of fit and tests for evaluating the statistical significance of the coefficients typically used in *PLS-PM* can be easily extended to the *QC-PM* approach (see Davino et al. 2016 for details).

4 Predictive-Oriented QC-PM

Exploiting the ability of QR to model the whole conditional distribution of a dependent variable, a Quantile Composite-based Path Modeling can be used in a prediction perspective. As above stated, the use of a dense grid of equally spaced quantiles provides an accurate approximation of the whole quantile process (Davino et al. 2013). The predictive model and the proposed procedure aims to provide the best in-sample predictions (often called fitted values). The computation of in-sample predictions requires estimating the parameters of a PLS model by using a given data sample and then using the model to predict values for cases of the same sample (Shmueli et al. 2016).

This paper proposes a two-step procedure to provide the best predictions of the outcome MVs:

1. *Estimation of several path models*

The first step aims to estimate the specified path model for a dense grid of equally spaced quantiles through the *QC-PM* algorithm. *QC-PM* provides m estimates for each parameter of the model, m LV scores and m previsions of the outcome MVs, where m is the number of the chosen quantiles. In particular, for the empirical analysis proposed in this paper, we exploited a grid of quantiles θ varying from 0.01 to 0.99 with a step of 0.01.

2. *Identification of the best model for each outcome MV and for each unit*

The second step aims to define the most predictive model for each outcome MV and the best accurate model for each statistical observation (henceforth *best quantile*). To achieve this goal it is necessary to compute the predictions corresponding to each quantile. Considering the outcome blocks, a partitioned table of predictions $\hat{\mathbf{X}}$ is obtained for each θ : $\hat{\mathbf{X}}(\theta) = [\hat{\mathbf{X}}_{J+Q+1}(\theta), \dots, \hat{\mathbf{X}}_K(\theta)]$. The generic element of a MV prediction, $\hat{\mathbf{x}}_{p_k}(\theta) = \{\hat{x}_{ip_k}(\theta)\}$, represents the prediction value of the MV \mathbf{x}_{p_k} , for the i -th unit according to the θ -th quantile.

The best model for each unit i and for each dependent MV of the k -th block, x_{p_k} , is identified by the quantile that best predict the variable, namely by the quantile which minimizes the absolute difference between the observed value and the estimated value:

$$\theta_{ip_k}^{best} = \operatorname{argmin}_{\theta=1, \dots, m} |x_{ip_k} - \hat{x}_{ip_k}(\theta)| \tag{8}$$

where $\theta_{ip_k}^{best}$ represents the quantile associated to the best predictive model for each unit i , each indicator p_k in each block k , while $\hat{x}_{ip_k}(\theta)$ is the correspondent best prediction for x_{ip_k} .

The denser the quantile grid is, the more accurate the forecasts provided by the predictive approach to *QC-PM* are. The best quantile provides an estimation of the unit position in the conditional distribution of the outcome variable. A comparison between the vector of the best quantiles for a given dependent variable and the corresponding unconditional

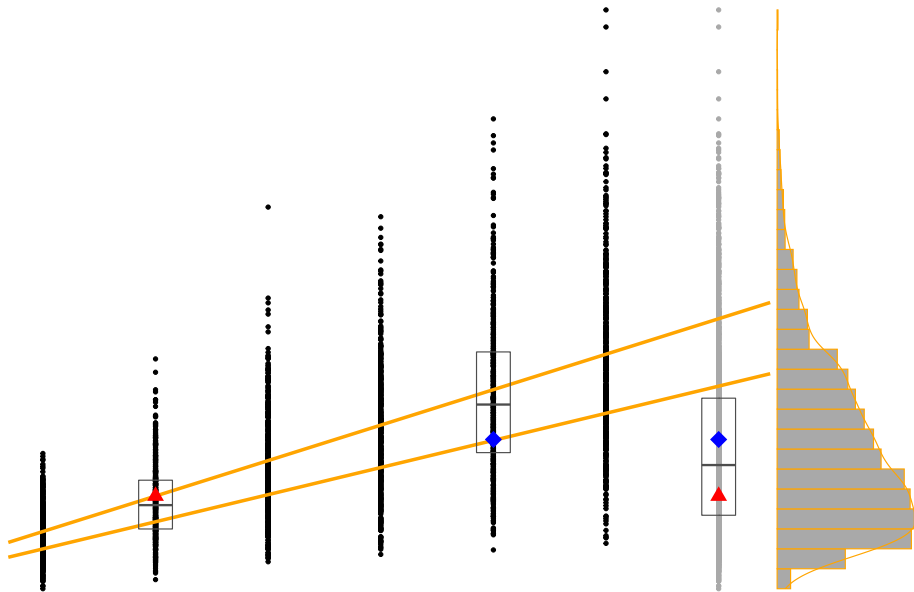


Fig. 2 An illustrative scatterplot to visualize the unconditional rank-quantiles (right grey dotplot) of two points (diamond and triangle) and their corresponding positions in the conditional distributions (second and fourth dotplot, from the left)

quantile (namely the position of each unit in the observed MV) allows to understand what is the effect of the structure of relationships in predicting each outcome variable.

A very simple example can be used to clarify the added value of the proposal and the different information that unconditional and conditional quantiles convey. Figure 2 shows the scatterplot of two variables observed on a sample of 3000 units. The dependent variable, represented on the vertical axis, has been generated from a Gamma model. Its unconditional distribution is represented through a boxplot, a dotplot and an histogram on the right-hand side of the figure using gray color. The regressor, represented on the horizontal axis, is a numerical variable with six values. The conditional distributions of the response on the six values of the regressor are depicted through the first six dotplots, starting from the left-hand side. The plot portrays a scenario in which the standard deviation exhibits a linear growth in the response variable as the regressor increases, the skewness and the excess kurtosis being instead constant and positive. Two observations are highlighted as an example: the diamond denotes an observation located above the median of the dependent variable distribution, with an unconditional quantile equal to 0.61; the triangle is placed in the lower tail, its unconditional quantile being 0.35. This is evident from the position of the two points in the gray right-hand boxplot. Starting from a dense grid of quantiles, the correspondent QR models were estimated, in order to identify the best quantile for each observation, namely the one that minimizes the function in Equation (8). The two lines depicted in Fig. 2 are the best QR models identified for the two example observations, respectively at the quantiles 0.31 for the diamond and 0.58 for the triangle. In this simple example, it is easy to prove that the identified best quantiles correspond to the estimation of the position of the two points in the conditional distributions. The two points are indeed highlighted according to their regressor value, namely in correspondence of the second and fifth conditional distribution (from left). They lie on the two best models whose θ values

are equivalent to their position in the conditional distribution. This is further highlighted superimposing a boxplot in order to assess their position.

5 An Empirical Analysis: The Prediction of Health Indicators

5.1 Data Description

This paper proposes an exploration of the relationships among *Health, Education and training* (henceforth EDU) and *Economic well-being* (henceforth ECO) using the “BES measures at local level” (ISTAT 2019a), a subset of the BES indicators, that are measured and regularly updated by Istat on the 110 Italian provinces and metropolitan cities (i.e. at NUTS3 level¹). The interest in such an application concerns both advances in knowledge about the dynamics producing the well-being outcomes at local level (multiplier effects or trade-offs) and a more complete evaluation of regional inequalities of well-being.

At the local level the well-being assets can strengthen each other affecting multiple disadvantages or advantages. For this reason it is important to consider both the levels and the relationships among BES indicators. Furthermore, equal results can be achieved in very different contexts and conditions. So, in assessing or comparing the well-being outcomes, the conditions within outcomes arise should be properly considered, adopting a multidimensional approach, able to support an accurate evaluation of the regional performances.

Figure 3 shows the specified network of relations. EDU, ECO and *Health* are the unobserved complex concepts that are measured as composites of the corresponding MVs (squares in Fig. 3 detailed in Tables 3 and 4 in the “Appendix”). Even though the model could be enriched by including further measures or domains, it still considers the most of the BES indicators that are currently produced by Istat at the NUTS3 level.

In the path model in Fig. 1, *Health* variables are placed as response variables of EDU and ECO. The underlying hypothesis, supported by literature and empirical studies (Costa et al. 2014; Mackenbach et al. 2008; Murtin et al. 2017; Petrelli et al. 2019), is that EDU has a direct effect on *Health* and an indirect effect mediated by ECO. In fact human capital is both a factor of economic competitiveness and well-being, as higher education offers more income opportunities, and promotes lower vulnerability to health risks.

With respect to the MVs, life expectancy at birth of males (O.1.1M) and females (O.1.1F) and infant mortality rate (O.1.2.MEAN_aa) are the three indicators used to measure the main global outcomes in the *Health* domain. EDU consists of indicators of qualification (O.2.2; O.2.3), competences (O_2.7_2.8; O_2.7_2.8_AA), participation in education and long-life learning (O.2.4; O.2.5aa; O.2.6). ECO is measured by indicators of income and wealth (O.4.1, O.4.2, O.4.3, O.4.5) and economic difficulties (O.4.4aa; O.4.6aa). All the indicators were positively oriented towards the BES (the higher is the indicator value, the greater is the BES) to provide an easier interpretation of the results. The “aa” suffix

¹ The acronym NUTS (from the French “Nomenclature des unités territoriales statistiques” NUTS) stands for Nomenclature of Territorial Units for Statistics, that is the European Statistical System official classification for the territorial units. The NUTS is a partitioning of the EU territory for statistical purposes based on local administrative units. The NUTS codes for Italy have three hierarchical levels: NUTS1 (Groups of regions); NUTS2 (Regions); NUTS3 (Provinces and Metropolitan Cities).

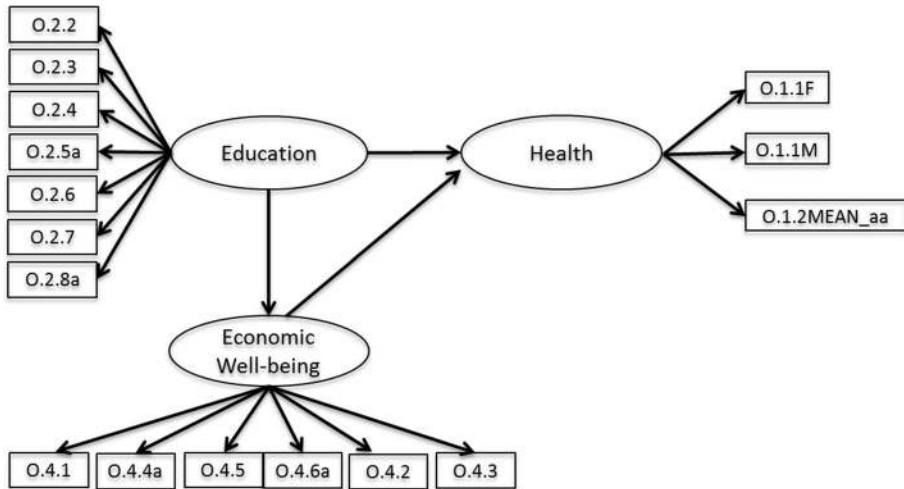


Fig. 3 The PLS-PM model

denotes those indicators that we reversed for this purpose². Data refer to the latest update available (reference year is given in the last column of Table 4 in the “Appendix”).

A preliminary analysis of the distribution of the MVs provides an examination of the heterogeneity that can be observed in the distribution of well-being in the Italian provinces. Figures 4, 5 and 6 show the violin plots, a combination of a box plot and a density plot. It is realized rotating and placing symmetrically on each side two density plots. The length of the vertical axis of each graph allows to appreciate the range of the observed values while the shape highlights how values are distributed in terms of variability and skewness. The black dot in the middle is the median value. Note that violin plots in different panels are not always comparable, as the variables have different unit of measurement and scales.

As expected, life expectancy at birth has a similar distribution for females (O.1.1F) and males (O.1.1M). However the median value is lower for men (80.5 years) compared to women (85.0). The gender gap (4.5 years comparing median values) is wide even looking at the ranges of the distributions: the maximum for men (82.1 years) is smaller than the minimum for women (82.8 years). In both cases the Italian provinces fall into ranges of equal width, but the male’s life expectancy has a more regular shape. Infant mortality is a rare phenomenon, so the corresponding MV (O.1.2 MEAN_aa) has a high territorial and temporal variability; for this reason the model was calculated on a three-year average. Even this more aggregate measure reveals large differences among the Italian provinces. The range is 5.1 points per thousand between the province with the worst result (equal to zero in the chart because the indicator was reversed) and the one with the best result. Most of the Italian provinces thicken in the centre of the distribution, with few cases placed in upper and lower ends. Therefore the major differences concern a small number of cases. The territorial heterogeneity in health outcomes does not have a clear geographical gradient. The provinces of central and northern Italy have more often better results than those

² To reverse the indicators we used the max-min method (OECD 2008).

Health block

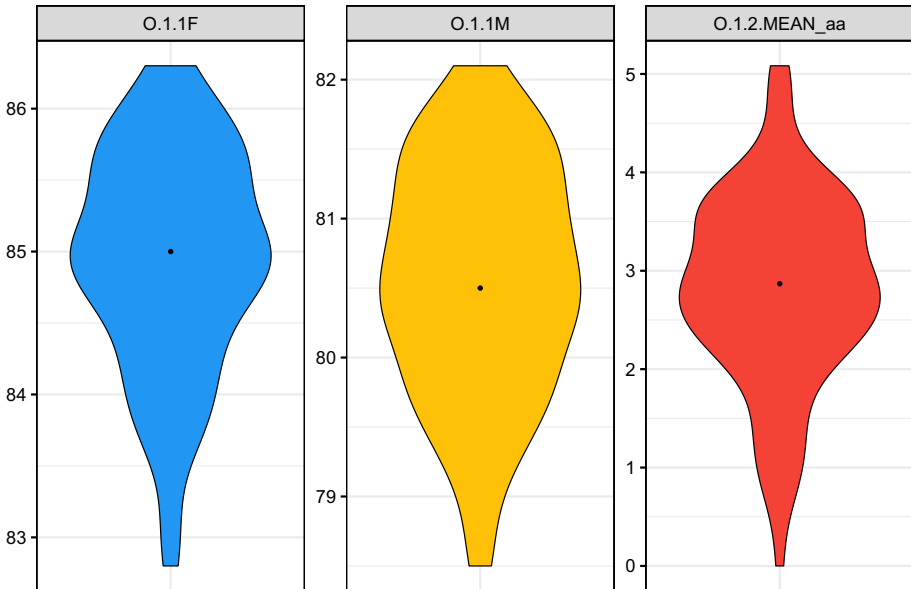


Fig. 4 Violin plot of the *Health* indicators

Education block

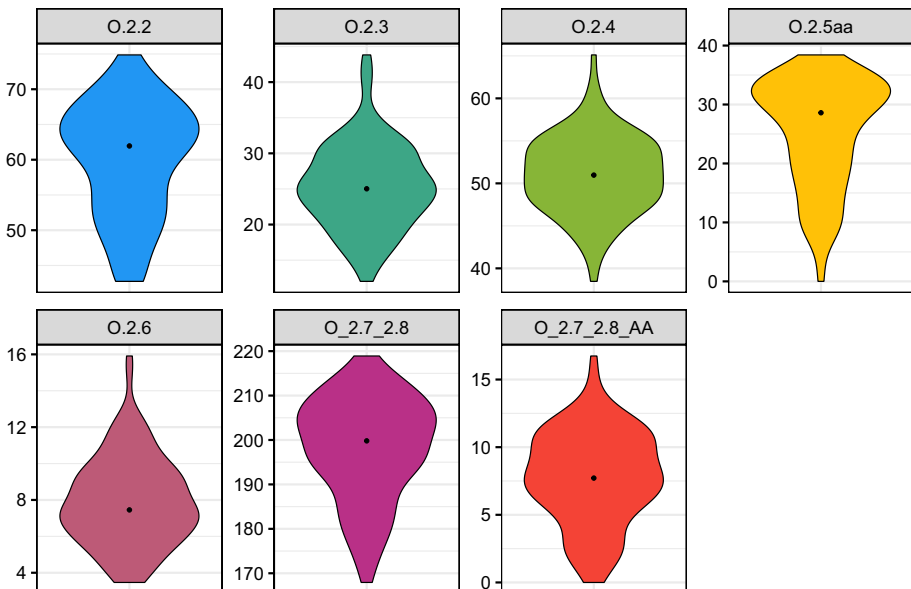


Fig. 5 Violin plot of the *Education* indicators

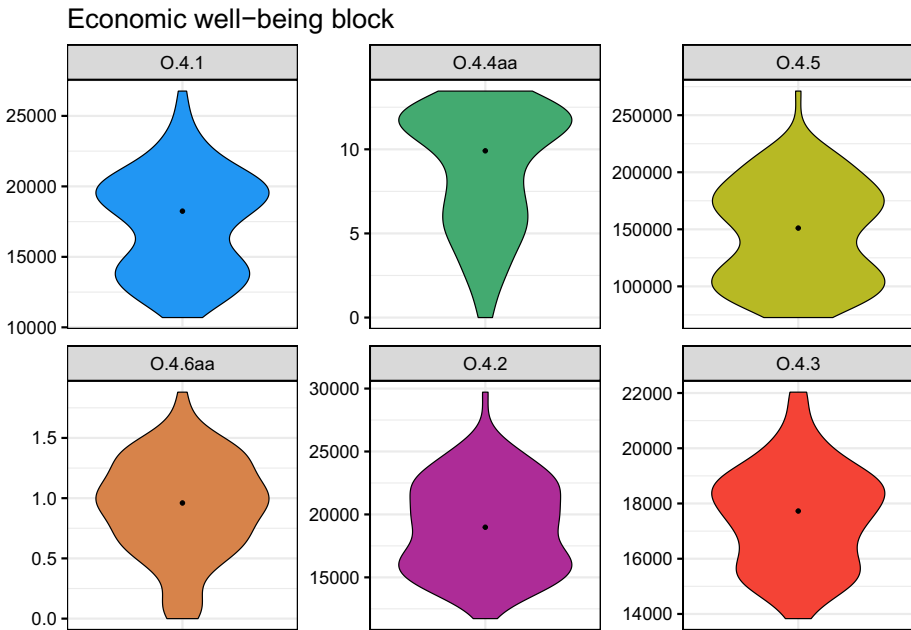


Fig. 6 Violin plot of the *Economic well-being* indicators

of the south and islands, with many positive and negative exceptions (ISTAT 2019b). Considering the *Health* predictors MVs, those of ECO are clearly the most discriminating. In particular all the indicators of income and wealth (O.4.1, O.4.2, O.4.3, O.4.5) have very asymmetric and polarized distributions, showing a sharp division between the group of provinces with the best economic outcomes and the group of the most penalized ones. The density of the median class is always lower compared to these two opposite groups. The shape of these graphs reflects a clear separation between the richer provinces of northern and central Italy, and the group of the southern and islands ones. The same division concerns the territorial distribution of low-income pensioners (O.4.4aa) which is very asymmetrical. In the Education block, the widest asymmetries emerge about the competences of young students (O_2.7_2.8) and the participation in education (O.2.5aa); both this measures oppose the southern Italian provinces, more disadvantaged, to the northern and central ones. Looking at the participation in lifelong learning (O.2.6) and at the highest qualification levels of the population (O.2.3) the territorial heterogeneity has a quite different sharp, that becomes thinner and longer moving towards the best outcomes, which therefore concern a few leading Italian provinces (namely the north-eastern ones for O.2.6 indicator and the northern metropolitan cities for O.2.3 indicator).

After examining the distribution of each MV, it is necessary to check the internal consistency of each block of MVs through the Cronbach's α and Dillon-Goldstein's ρ , which need to be greater than 0.7. For the Cronbach's α , Confidence Intervals are also reported following a recent approach (Trinchera et al. 2018). Moreover, the average variance extracted (AVE) is also considered (Tenenhaus et al. 2005). Table 1 shows that, for all the blocks, internal consistency is satisfied and all the AVE values are greater than 0.5.

Table 1 Reliability and internal consistency measures

	Cronbach's α (95% C.I.)	Dillon–Goldstein's ρ	AVE
<i>Education</i>	0.907 (0.885; 0.928)	0.927	0.647
<i>Economic well-being</i>	0.953 (0.942; 0.964)	0.963	0.813
<i>Health</i>	0.733 (0.640; 0.827)	0.853	0.666

5.2 PLS-PM and QC-PM Results

The model in Fig. 3 was estimated using the classical *PLS-PM* and *QC-PM* fixed to the three quartiles ($\theta=[0.25, 0.5, 0.75]$). Given the territorial heterogeneity expressed by the model MVs, the aim is to explore whether estimates vary across different parts of the variable distributions.

Firstly, the estimated loadings in the measurement models are examined. Table 2 presents the loadings for each MV estimated using conventional *PLS-PM* and *QC-PM* at the three defined quartiles.

Except for few cases, loadings are very similar across quartiles and compared to the one estimated by *PLS-PM*. However, for each block of MVs, measurement invariance should be verified to evaluate any potential difference in the structural relationships across the various quartiles (i.e., comparison among path coefficients). *QC-PM* still lacks a statistical test for measurement invariance, but we applied the variant of *QC-PM* fixing the quantile in the measurement model to the median and found no relevant differences in results, hence the measurement of the LV remains essentially the same for all quartiles, which allows a reliable comparison of path coefficients across the various quartiles.

Bars in each panel of Fig. 7 represent (from the top to the bottom) the path coefficients and the standard errors measuring respectively the effects on the conditional average and on the conditional quartiles of *Health*. It is interesting to note how *QC-PM* results complement *PLS-PM* results. Looking at the average of the distribution (*PLS-PM*), *Education* is the most important driver of *Health*, but *QC-PM* reveals that its importance is greater where the *Health* is lower (that is lower or equal to the median) while it decreases as the *Health* grows. In essence, the effect of *EDU* on health conditions increases moving from provinces with good to worse results. With regard to *ECO*, the additional information provided by *QC-PM* is also interesting because the *PLS-PM* results suggests that *ECO* does not contribute to *Health* while the *QC-PM* reveals an high path coefficient in those provinces where *Health* scores are the highest ever. Therefore *ECO* plays a discriminating role in explaining the best absolute outcomes.

The territorial heterogeneity of the MVs is often associated with geographical differences; therefore it may be useful to add the geographical location of the provinces in the analysis of the model results (Davino et al. 2017). A possible source of heterogeneity could be, for example, the geographical area considering that Italian provinces are usually grouped into four areas: north-east (20%), north-west (23%), centre (20%) and south and islands (37%). Figure 8 shows all the possible scatter plot combining the three composites obtained by estimating the model in Fig. 3 with *PLS-PM*. In each panel a different color and shape of the points is used to distinguish the effect of the area. The lines represent the regression lines estimated in the four subgroups of provinces considering the variable represented on the vertical axis as dependent variable. The boxplots on the right (top) side of each panel show the distribution of the composites represented on the vertical (horizontal)

Table 2 Measurement model results

	Loadings (λ)	PLSPM	Estimation at different quantiles		
			$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$
Health	<i>O.1.1.F</i>	0.93	0.98	0.91	0.94
	<i>O.1.1.M</i>	0.89	0.92	0.88	0.85
	<i>O.1.1.2.Mean_aa</i>	0.57	0.65	0.61	0.29
Education	<i>O.2.2</i>	0.86	0.93	0.89	0.82
	<i>O.2.3</i>	0.85	0.81	0.79	0.77
	<i>O.2.4</i>	0.63	0.61	0.62	0.65
	<i>O.2.5.aa</i>	0.91	0.90	0.96	0.94
	<i>O.2.6</i>	0.71	0.63	0.81	0.82
	<i>O._2.7_2.8</i>	0.85	0.76	0.85	0.85
	<i>O._2.7_2.8_AA</i>	0.78	0.78	0.77	0.75
Economic Well-Being	<i>O.4.1</i>	0.98	0.93	0.97	1.03
	<i>O.4.4aa</i>	0.90	0.91	0.92	0.90
	<i>O.4.5</i>	0.93	0.85	0.91	1.02
	<i>O.4.6aa</i>	0.76	0.85	0.68	0.75
	<i>O.4.2</i>	0.92	0.92	0.92	0.96
	<i>O.4.3</i>	0.91	0.87	0.88	0.89

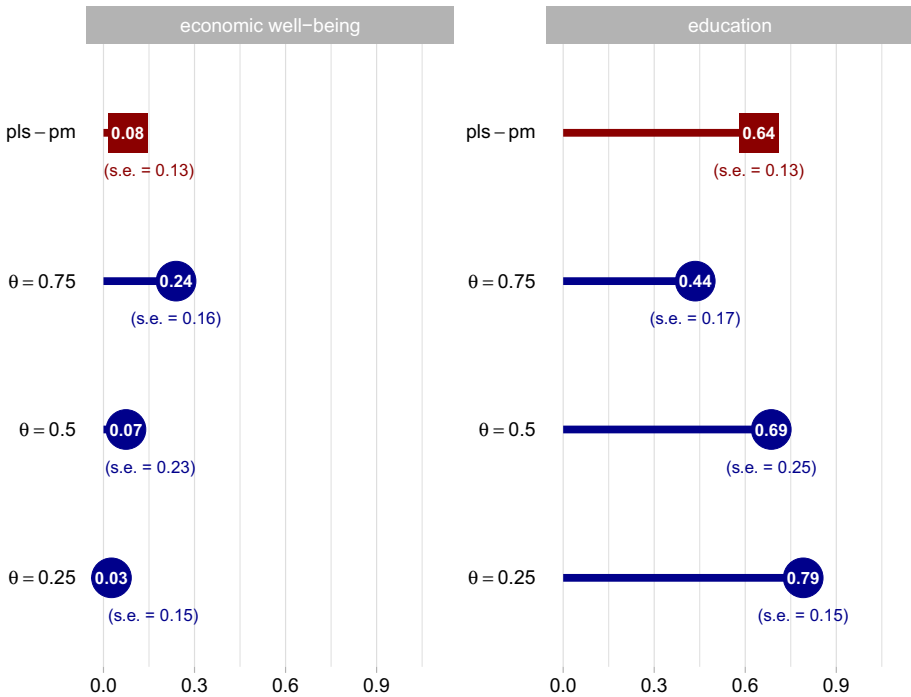


Fig. 7 Path coefficients linking Economic well-being and Education to Health

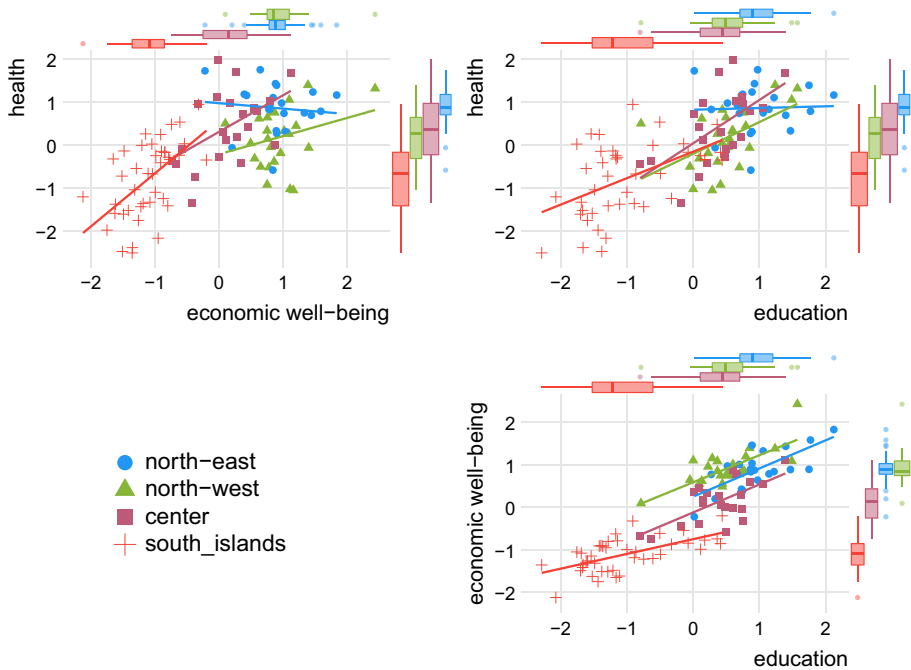


Fig. 8 Education, Economic well-being and Health distributions according to the geographic area. PLS-PM results

axis distinguishing once again the provinces by geographical area. Considering the univariate distributions of the composites, it is clear that the southern provinces are lagging behind in all the three contexts analysed; the third quartile for the southern provinces is always well below the first quartile for the other provinces. The gap is broader for EDU and ECO.

As expected, given the distribution of EDU and ECO MVs (Sect. 5.1), the composites distribution has a clear geographical orientation: values increase moving from the south to the north of Italy, with the north-east group leading. Conversely, the *Health* distribution follows a different geographical progression. The group of central Italian provinces, although very heterogeneous, tends to have better scores than all the other groups, including the north-eastern one.

Considering the relationships among *Health*, ECO and EDU, the three composites are highly correlated at a national level but differently at local level. In Fig. 8, for each couple of scores, a regression line is estimated in the four geographical areas and superimposed to the scatter plot. The simultaneous representation of the scatter plot and regression lines allows to capture both the trend and the heterogeneity of the relationship. The correlation between the ECO and EDU composites is equally strong in all the four geographical areas despite the heterogeneity observed within each group of provinces (Fig. 8, bottom right panel). The correlation between *Health* and ECO it is by far the strongest in the south and islands group, according to the greater path coefficient of ECO on the *Health* worst outcomes (note that the south and islands provinces always lie in the *Health* distribution queue, with just one exception). Instead the strongest

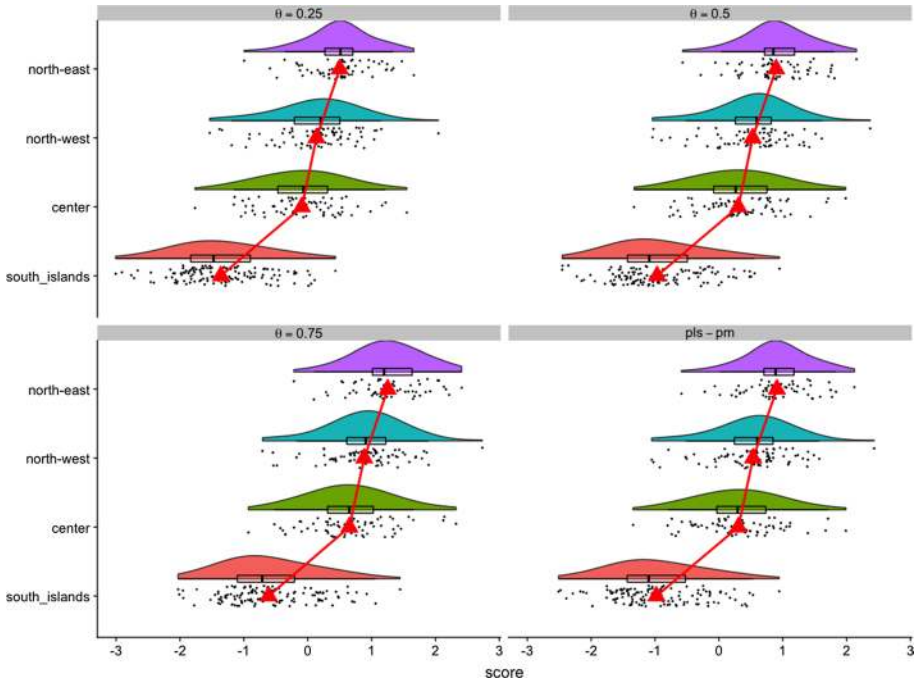
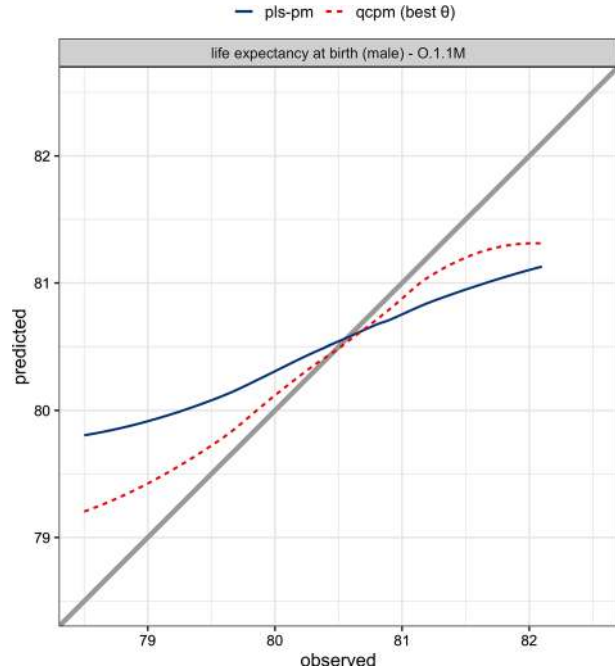


Fig. 9 Distribution of the Health composite from a PLS-PM (top in each panel) and *QC-PM* estimated at the three quartiles, according to the geographic area

correlation between *Health* and EDU scores is that of the center group. Looking at the *PLS-PM* results shown in Fig. 8, the assumptions underlying the model are not confirmed for all the north-eastern provinces: also due to the high dispersion and heterogeneity of this provinces, no correlation arise between *Health* and EDU outcomes, while that between *Health* and ECO (very weak) still has a negative sign.

The results of the *QC-PM* can provide a better definition of the characteristics of this heterogeneity. Focusing on the *Health* composite and on its three conditional quartiles, it is possible to analyse similarities and differences among the geographical areas at different health conditions. Figure 9 shows the distribution of the *Health* composite for each area (different panels) and for each model (rows in each panel). The density plot, the dot diagram and the boxplots allow to explore all the features of the distributions. In each line a segment joins the averages of the composite at the three quartiles. Considering that the global averages of the composites provided by the *PLS-PM* and by *QC-PM* at each quartile are equal respectively to 0, -0.47, 0.01 and 0.47, it is possible to note that the averages of the southern provinces distributions are always below the global average, while north-eastern provinces (and partially also the north-western ones) show an opposite behavior.

Fig. 10 A smoothed version (using a linear smoother) of the scatterplot of observed and predicted values for the O.1.1M variable obtained using a *PLS-PM* (solid line) and the best quantile approach (dashed line)



5.3 Prediction Results

In order to exploit the model in Fig. 3 in an operative predictive perspective, it is necessary to consider that *Health* is composed by three outcome variables: life expectancy at birth of males (O.1.1M) and females (O.1.1F) and infant mortality rate (O.1.2.MEAN_aa). In deepening the presentation of the prediction results, we will just focus on O.1.1M, which, as seen above (Sect. 5.1) is the most informative among the three *Health* indicators we used in the model. In fact it is more robust than the infant mortality rate (that is affected by extra-variability) and, compared to females rates, it is more able to explain both the differences among Italian provinces and the improvements achieved over time in the general level of the total life expectancy in Italy. As we deal with in-sample prediction, the post-analysis of the results can be based on the comparison between the observed and the prediction values. As described in the methodological sections, the proposed best quantile approach computes the predicted vector of each MV by selecting the quantile model providing the best prediction for each province. In case of a simple regression model (with one dependent variable and one regressor) the identification of the best quantile allows to exactly reconstruct the observed variable. In a more complex model as the network of relationships in Fig. 3 is, the goal is to identify the best prediction.

Figure 10 shows a smoothed version (using a linear smoother) of the scatterplot of observed and predicted values for the O.1.1M variable where the predicted values derives from a *PLS-PM* (solid line) and from the best quantile approach (dashed line). The gray line depicts the bisector, namely the place of the points where the observed values and the expected values coincide perfectly. It results that the best quantile predictions are much more accurate than *PLS-PM*, but the marginal gain in accuracy decreases at the distribution tails. This consideration is also confirmed in the analysis by geographical area

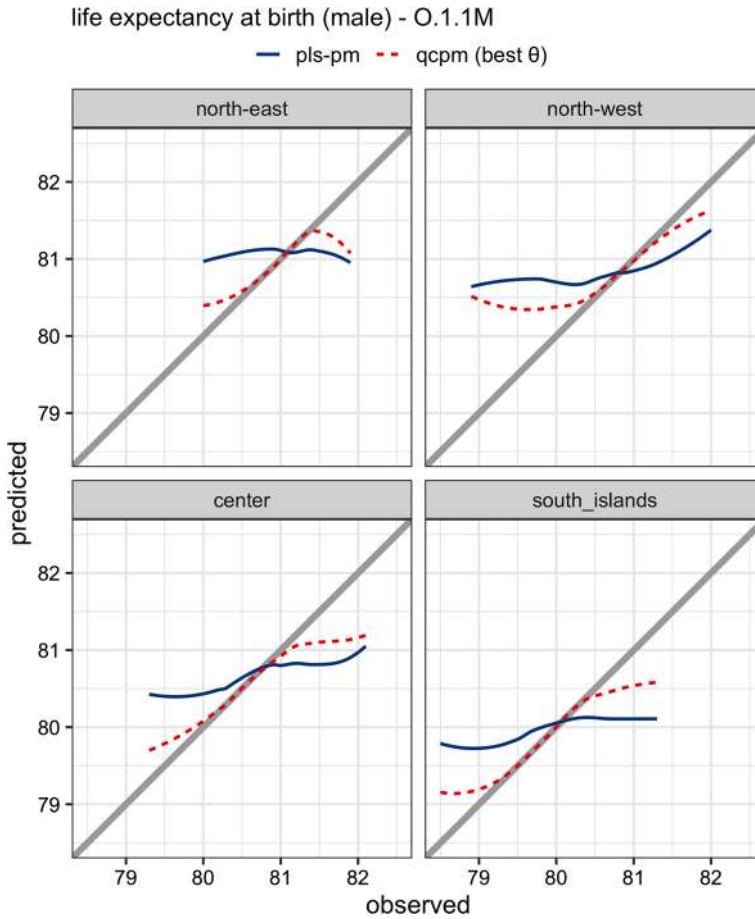


Fig. 11 A smoothed version (using a linear smoother) of the scatterplot of observed and predicted values for the O.1.1M variable obtained using a *PLS-PM* (solid line) and the best quantile approach (dashed line) according to the geographic area

(Fig. 11): the higher accuracy of the best quantile predictions is particularly evident for the north-east and south and islands areas (the dashed lines are very closed to the bisector) while the lower accuracy of the best quantile predictions in the extreme parts of the distribution is more evident in north-west (low tail) and center (high tail).

The proposed quantile model-based prediction approach can provide useful information to understand, for each province, what is the contribution of the system of relationships in the model in Fig. 3 to the prediction of health levels. In essence, the comparison between conditional and unconditional quantiles for the *Health* MVs tells us if the observed results are in line with the starting conditions (in terms of ECO and EDU).

The scatter plot in Fig. 12 visualizes all the provinces according to the assigned best quantiles and to the unconditional quantiles. The unconditional quantile (horizontal axis) is the position of each province in the MV distribution without considering/controlling the effect played by EDU and ECO, while the best quantile (vertical axis) represents the

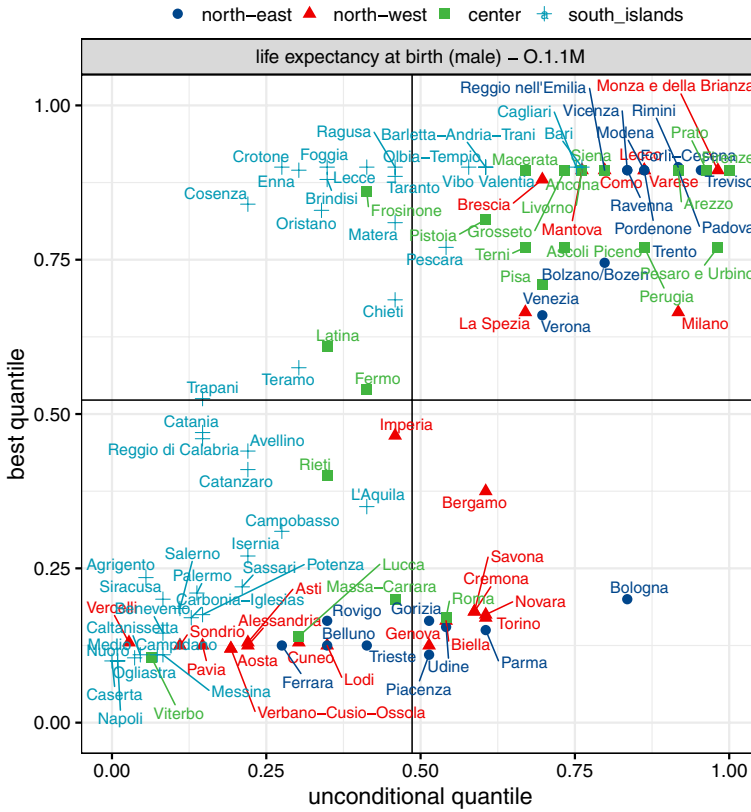


Fig. 12 The scatter plot of the provinces according to the unconditional and conditional best quantiles of O.1.1M. The color and shape of the points represent the geographic area

position of each province compared to all the other provinces that have similar EDU and ECO levels (i.e., the position in the conditional distribution).

If on one hand, looking at the unconditional quantiles of O.1.1M one can find vertically aligned provinces that have the same value of the life expectancy of males also starting from different EDU and ECO levels, on the other hand looking at the horizontal alignment of the points, one can find provinces that perform similarly while having different socioeconomic levels. The furthest points from the bisector of the scatterplot identify the territories where the differences between observed and “potential” results are greatest. These divergences are the most interesting cases to study. By numbering the quadrants counterclockwise and starting from the quadrant at the top right, we can identify two “critical” situations: at the top left (second quadrant) fall those provinces that get better results than the expected ones (the best quantile is greater than the unconditional) and at the bottom right (fourth quadrant) fall those territories with an important negative gap (the best quantile is much lower than the observed one). The geographical area has some influence on the relationship between unconditional and conditional quantiles. In fact, in the second quadrant we find almost exclusively southern provinces (there are only Latina, Frosinone and Fermo for the center) while in the fourth quadrant we find only northern provinces together with Rome. The scatter

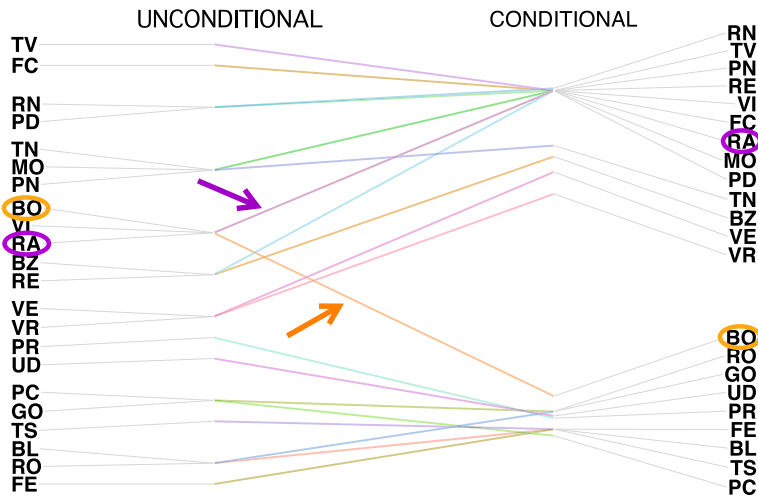


Fig. 13 The slope graph of the sub-group of provinces in the north-east of Italy. Provinces are ranked according to their position in the original (unconditional, left-hand side) and estimated (conditional, left-hand side) distribution of O.1.1M

confirms the advantage of the center and northern areas and the penalisation of the south and islands also for the males life expectancy at birth. However in this multidimensional perspective the dimensions of the advantages and disadvantages are quite different from those we could appreciate in a context of univariate analysis (i.e. considering the single indicators and not also their interrelationships).

For a more analytical illustration of the additional information that the model provides, an extreme example can be isolated: Bologna vs Ravenna. Bologna and Ravenna share quite similar health outcomes, considering the O.1.1M variable, but the predicted values are quite different as shown by their positions respectively in the fourth quadrant (Bologna) and in the second one (Ravenna). If we consider the subset of the provinces of the north-eastern area, it is more evident the effect induced by the model in the *Health* results of the two provinces. The slope graph in Fig. 13 shows the sub-group of provinces in the north-east of Italy ranked according to their position in the original (unconditional, left-hand side) and estimated (conditional, left-hand side) distribution of O.1.1M. The slope of the lines joining the unconditional and conditional position of each province clearly visualize how much taking into account the levels of EDU and ECO can affect the life expectancy of males. The limit case is represented by a horizontal line: it would mean that ECO and EDU levels make no contribution to the knowledge of *Health*. Both for Bologna and Ravenna, the results in terms of life expectancy of males, estimated in itself, are excellent, among the highest: the two provinces share the 83-th percentile in Italy. However, the effects of the estimated model are different in the two provinces resulting in an improvement in the position occupied by Ravenna (increasing slope of the stick) and a worsening for Bologna (decreasing slope of the stick), the two best quantiles being 0.20 and 0.89, respectively.

To interpret the different effect played by the model, with similar observed results, it is necessary to go back to the distribution of the original indicators. Figure 14 and 15 show some univariate statistics of the ECO and EDU through parallel coordinates. The grey double lines join the quartiles of the indicators while the thin lines represent the averages by

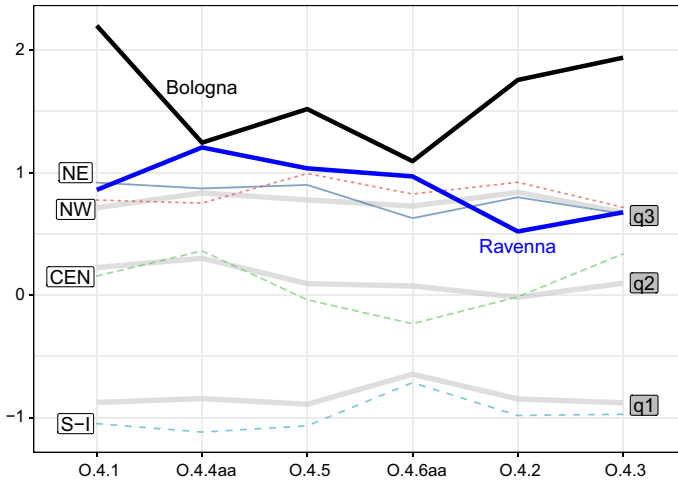


Fig. 14 Distribution of the univariate statistics of the Economic well-being indicators: quartiles (grey double lines), averages by geographical area (thin lines). The broken lines representing Ravenna and Bologna are highlighted

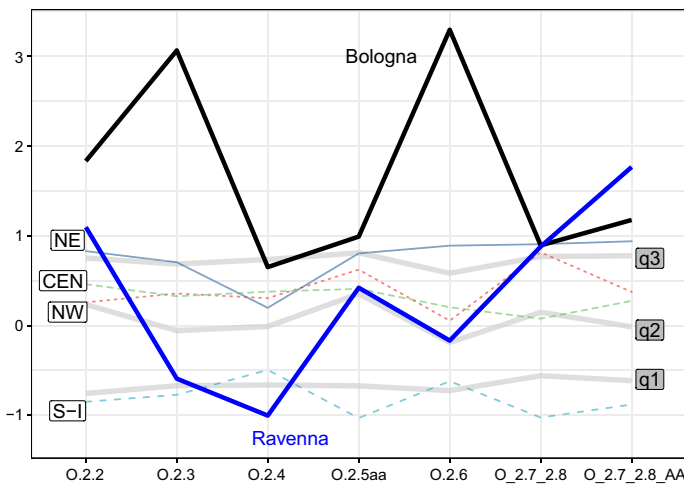


Fig. 15 Distribution of the univariate statistics of the Education indicators: quartiles (grey double lines), averages by geographical area (thin lines). The broken lines representing Ravenna and Bologna are highlighted

geographical area. The broken lines representing Ravenna and Bologna are highlighted and allow to contextualize why the performance of Bologna appears less brilliant. If on one hand the levels of EDU and ECO in Bologna are among the highest, on the other hand in the group of provinces with these highest levels of EDU and ECO, Bologna ranks among the last in terms of life expectancy of males, as highlighted by the best quantile that is much lower than the unconditional one. This difference gives us a measure of the gap between potential and actual results, which in this case is negative. Reading the results

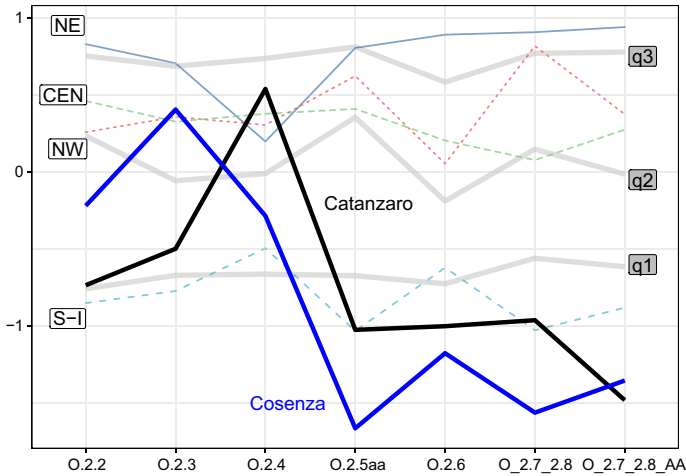


Fig. 16 Distribution of the univariate statistics of the Education indicators: quartiles (grey double lines), averages by geographical area (thin lines). The broken lines representing Catanzaro and Cosenza are highlighted

from the point of view of Ravenna, it appears that Ravenna has the same life expectancy of males as Bologna (81.5 years), but a much less favorable ECO context, especially concerning the disposable income of households (O.4.1), the incomes of employees (O.4.2), pensioners (O.4.3) and, the wealth of families (O.4.5), and it is shown above that in the model ECO has a higher path coefficient on the higher part of the Health distribution. The gaps in terms of EDU are even more marked in particular on the rates of graduates (O.2.3), transition to university (O.2.4) and participation in lifelong learning (O.2.6); moreover we know that EDU in the model has a high path coefficient on the higher part of the Health distribution. Given its advantage over Ravenna in terms of EDU and ECO, Bologna should have a far better result in terms life expectancy of males than what is observed.

An opposite pattern can be exemplified by Cosenza and Catanzaro (Fig. 16). Cosenza falls in the group of provinces with low results (81st with 79.9 years), but gets a better position than expected, given ECO and EDU, as its conditional quantile is higher than the unconditional and the positive gap is quite wide. So we could say that Cosenza performs better than Catanzaro, as it gets the same result but in a more unfavorable context, mostly due to the lower levels of the indicators of ECO (see Fig. 17).

6 Conclusions and Further Developments

The analysis of the relationships among complex and unobservable factors can be enhanced using a quantile approach to *PLS-PM*, which allows to highlight the unobserved heterogeneity that could be overlooked by the classic estimation of the average effects. In this paper *QC-PM* is also proposed in a predictive perspective providing the best estimation, and thus the best model, associated to each statistical unit. For a given statistical unit, the quantile associated to the best model, in the paper named "best quantile", condenses the effect played by the regressors on the position of the unit in the conditional distribution of the dependent variable. *QC-PM* lacks a statistical test for measurement invariance,

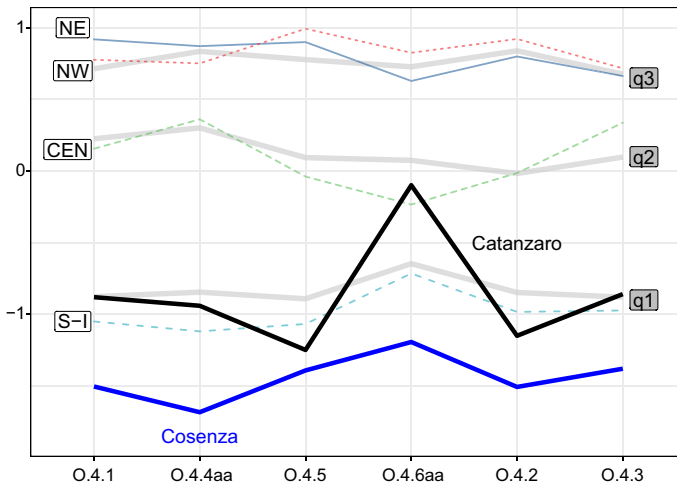


Fig. 17 Distribution of the univariate statistics of the Economic well-being indicators: quartiles (grey double lines), averages by geographical area (thin lines). The broken lines representing Catanzaro and Cosenza are highlighted

which allows a reliable comparison among path coefficients estimates over quantiles. However, as above-mentioned, a possible variant of QC-PM can be used, fixing the quantile in the measurement model to the median and changing only the quantile in the structural model. We applied this variant, but we did not find relevant differences in results, probably because there are not evident differences among loadings over quantiles.

The potential arising from a joint use of a *PLS-PM* and a *QC-PM* are exploited to explore the relationships among the *Health* outcomes and the levels of *Economic well-being* and *Education* in Italian provinces. The model is defined using a subset of dimensions and indicators of the well-known BES dataset produced by ISTAT at NUTS3 level (ISTAT 2019a). The underlying idea is that health levels and health inequalities at local level can be assessed more in depth taking into account both the observed and the unobserved heterogeneity. In fact similar levels of health can result from very different performances, when they are achieved in different socio-economic conditions. The study provided a multidimensional analysis of health inequalities at local level, in the effort to capture the unobserved heterogeneity that can be explained taking into account the relationships among *Health*, *Economic well-being* and *Education*.

The results of the *PLS-PM* confirmed that there is a relationship between *Education* and *Health*, as we hypothesized in the theoretical model. The *QC-PM* also revealed the existence of a relevant relationship between *Economic well-being* and high levels of *Health* and a decreasing impact and contribution of *Education* to increasing levels of *Health*. The geographical area also provided useful information for understanding differences in *Health* levels and in the relations among *Health*, *Economic well-being* and *Education*. Globally, the *PLS-PM* confirmed that the three well-being domains are highly correlated in all geographical areas, except the north-east. Deepening the analysis in a predictive perspective the best quantile predictions resulted much more efficient than *PLS-PM*, especially concerning the north-east subgroup. The observed health results of each province could then be assessed taking into account jointly its placement in the unconditional distribution, the results of the best quantile prediction and its geographical location. Looking at life

expectancy of males, many provinces, almost all located in the south and islands, get low but better results than expected; in contrast, the provinces that get high but lower results than expected are less numerous and none of them is southern or insular.

Future research will explore the strengths of QC-PM for prediction outside the data sample used for estimating the model (out-of-sample prediction) and will try to deepen the knowledge about the determinants of health differences at the local level, including in the model the relationships among health and other well-being assets, such as the environment, the quality of health services, the exposure to risky jobs and other vulnerability factors.

Acknowledgements Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Compliance with ethical standards

Conflict of interest The authors declare that they have no relevant or material financial interests that relate to the research described in the paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

See Tables 3, 4.

Table 3 Manifest variables for the Health block and the Education block

Domain	Label	Indicator (MV)	Description	Unit of measurement	Reference year
Health	<i>O.1.1.F</i>	Life expectancy at birth (females)	Life expectancy expresses the average number of years that a child born in a given calendar year can expect to live if exposed during his whole life to the risks of death observed in the same year at different age	Average number of years	2017
	<i>O.1.1.M</i>	Life expectancy at birth (males)			
Education	<i>O.1.2.MEAN_aa</i>	Infant mortality rate	Ratio of children dead during the first year of life to the total number of children born in the same year	Per 1.000 born alive (3 years mean)	2014-2016
	<i>O.2.2</i>	People with at least upper secondary education level (25-64 years old)	Ratio of people aged 25-64 years having completed at least upper secondary education (ISCED [1] level not below 3) to the total of people aged 25-64 years, [1] ISCED is the UNESCO International Standard Classification of Education for degree programs and related degrees. Level 3 is the Upper secondary education degree, Level 5 is the First stage of tertiary education degree.	Percentage	2018
	<i>O.2.3</i>	People having completed tertiary education (30-34 years old)	Ratio of people aged 30-34 years having completed tertiary education (ISCED 5, 6, 7 or 8) to the total of people aged 30-34 years.	Percentage	2018
<i>O.2.4</i>	First-time entry rate to university by cohort of upper secondary graduates	Proportion of new-graduates from upper secondary education enrolled for the first time at university in the same year of upper secondary graduation		Cohort-specific percentage rate	2017
<i>O.2.5.aa</i>	People not in education, employment or training (Neet)		Ratio of people aged 15-29 years that are not in education, employment, or training to the total people aged 15-29 years	Percentage	2018

Table 3 (continued)

Domain	Label	Indicator (MV)	Description	Unit of measurement	Reference year
	O_2.6	Participation in long-life learning	Ratio of people aged 25-64 years participating in formal or non-formal education to the total people aged 25-64 years	Percentage	2018
	O_2.7_2.8	Level of literacy and numeracy	Scores obtained in the tests of functional skills of the students in the II classes of upper secondary education	Average score on a 0-200 scale	2018
	O_2.7_2.8_AA	Gender differences in the level of numeracy and literacy	Differences between males and females students in the level of numeracy and literacy	Absolute difference between specific average scores	2018

Table 4 Manifest variables for the Well-being block

Domain	Label	Indicator (MV)	Description	Unit of measurement	Reference year
Well Being	<i>O.4.1</i>	Per capita disposable income	Ratio of total disposable income of households to the total number of residents	Euro	2016
	<i>O.4.4aa</i>	Pensioners with low pension amount	Pensioners who receive a monthly gross pension of less than 500 Euros to the total number of pensioners	Percentage	2017
	<i>O.4.5</i>	Per capita net wealth	Ratio of total net wealth of households to the total number of residents	Thousands of euro	2016
	<i>O.4.6aa</i>	Rate of bad debts of the bank loans to families	Ratio of the amounts of new non-performing loans in the year (loans to subjects declared insolvent or difficult to recover during the year) to the total stock of non-performing loans during the year	Percentage	2017
	<i>O.4.2</i>	Average annual salary of employees	Ratio of the total annual remuneration (gross of personal income tax) of non-agricultural private sectors employees to the number of employees	Euro	2017
	<i>O.4.3</i>	Average annual amount of pension income per capita	Ratio of the total amount of pensions paid in the year to the total number of pensioners	Euro	2017

References

- Costa, G., Bassi, M., Gensini, G. F., Marra, M., Nicelli, A. L., & Zengarini, N. (2014). *L'equità nella salute in Italia. Secondo rapporto sulle disuguaglianze sociali in sanità*. Roma: Franco Angeli.
- Danks, N., & Ray, S. (2018). Predictions from partial least squares models. In F. Ali, S. Rasoolimanesh, & C. Cobanoglu (Eds.), *Applying partial least squares in tourism and hospitality research* (pp. 35–52). Emerald Publishing Limited.
- Davino, C., Esposito Vinzi, V., & Dolce, P. (2016). Assessment and validation in quantile composite-based path modeling. In H. Abdi, V. Esposito Vinzi, G. Russolillo, G. Saporta, & L. Trinchera, (Eds.), *The multiple facets of partial least squares and related methods* (pp. 169–185). Springer Proceedings in Mathematics & Statistics. New York: Springer.
- Davino, C., Dolce, P., & Taralli, S. (2017). Quantile composite-based model: A recent advance in pls-pm. a preliminary approach to handle heterogeneity in the measurement of equitable and sustainable well-being. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: basic concepts, methodological issues and applications* (pp. 81–108). Cham: Springer.
- Davino, C., Dolce, P., Taralli, S., & Esposito Vinzi, V. (2018). A quantile composite-indicator approach for the measurement of equitable and sustainable well-being: A case study of the Italian provinces. *Social Indicators Research*, 136, pp. 999–1029, Dordrecht, Kluwer Academic Publishers.
- Davino, C., & Esposito Vinzi, V. (2016). Quantile composite-based path modelling. *Advances in Data Analysis and Classification*, 10(4), 491–520.
- Davino, C., Furno, M., & Vistocco, D. (2013). *Quantile regression: Theory and applications*. Wiley, Wiley Series in Probability and Statistics.
- Davino, C., & Vistocco, D. (2015). Quantile regression for clustering and modeling data. In I. Morlini, T. Minerva, & M. Vichi (Eds.), *Advances in statistical models for data analysis: studies in classification, data analysis, and knowledge organization* (pp. 85–96). Heidelberg: Springer.
- Davino, C., & Vistocco, D. (2018). Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on University outcome. *Statistics & Its Interface*, 11, 541–556.
- Di Napoli, I., Dolce, P., & Arcidiacono, C. (2019). Community trust: A social indicator related to community engagement. *Social Indicators Research*, 145(2), 551–579.
- Dolce, P., Esposito Vinzi, V., & Lauro, C. N. (2017). Predictive path modeling through PLS and other component-based approaches: Methodological issues and performance evaluation. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 153–172). Cham: Springer.
- Dolce, P., & Hanafi, M. (2017). Multi-dimensional blocks in predictive path modeling, 9th international conference on pls and related methods (PLS'17), Macau, China, 17–19 June 2017.
- Dolce, P., Esposito Vinzi, V., & Lauro, C. (2018). Non-symmetrical composite-based path modeling. *Advances in Data Analysis and Classification*, 12(3), 759–784.
- Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of partial least squares*. Springer.
- Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. *Annals of Mathematical Statistics*, 35(3), 1019–1030.
- Fox, M., & Rubin, H. (1964). Admissibility of quantile estimates of a single location parameter. *Journal of Business Research*, 69(10), 4565–4582.
- Furno, M., & Vistocco, D. (2018). *Quantile regression: Estimation and simulation*. Wiley, Wiley Series in Probability and Statistics.
- Glang, M. (1988). *Maximierung der Summe erklärter Varianzen in linearrekursiven Strukturgleichungsmodellen mit multiple Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahren*. Phd thesis, Universität Hamburg, Hamburg, Germany.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Thousand Oaks, CA: Sage.
- Hair, J., Sarstedt, M., Pieper, T., & Ringle, C. (2012). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, 45, 320–340.
- Hanafi, M. (2007). PLS path modeling: Computation of latent variables with the estimation mode B. *Computational Statistics*, 22, 275–292.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, 33(3), 405–431.
- Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 277–319.

- ISTAT. (2013). Rapporto Bes 2013. Il benessere equo e sostenibile in Italia. Roma, Istat. <https://www.istat.it/it/archivio/84348>.
- ISTAT. (2018). Bes report 2018: Equitable and sustainable well-being in Italy. Rome, Istat. <https://www.istat.it/en/archivio/225140>.
- ISTAT. (2019a). Misure del Benessere dei territori. Tavole di dati. Rome, Istat. <https://www.istat.it/it/archivio/230627>.
- ISTAT. (2019b). Le differenze territoriali di benessere - Una lettura a livello provinciale. Rome, Istat. <https://www.istat.it/it/archivio/233243>.
- Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.
- Koenker, R., & Basset, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Krämer, N. (2007). *Analysis of high-dimensional data with partial least squares and boosting*. Phd thesis, Technische Universität Berlin, Berlin, Germany.
- Lohmöller, J. B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg: Physica-Verlag.
- Mackenbach, J. P., Stirbu, I., Roskam, A. J., Schaap, M. M., Menvielle, G., Leinsalu, M., et al. (2008). European union working group on socioeconomic inequalities in health. Socioeconomic inequalities in health in 22 European countries. *The New England Journal of Medicine*, 358, 2468–2481.
- Mathes, H. (1993). Global optimisation criteria of the PLS-algorithm in recursive path models with latent variables. In K. Haagen, D. Bartholomew, & M. Deister (Eds.), *Statistical modelling and latent variables*. Amsterdam: Elsevier Science.
- Murtin, F., Mackenbach, J., Jasilionis, D., & Mira d’Ercole, M. (2017). Inequalities in longevity by education in OECD countries: Insights from new OECD estimates”, *OECD Statistics Working Papers*, No. 2017/02, OECD Publishing, Paris.
- OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. Paris: OECD.
- Petrelli, A., Di Napoli, A., Sebastiani, G., Rossi, A., Rossi, P. Giorgi, Demuru, E., Costa, G., Zengarini, N., Alicandro, G., Marchetti, S., Marmot, M., & Frova, L. (2019). Italian Atlas of mortality inequalities by education level. *Epidemiologia e prevenzione*, 43, 1S1: 1–120.
- Sarstedt, M., Ringle, C. M., & Hair, J. F. (2017). Partial least squares structural equation modeling. In C. Homburg et al. (Eds.), *Handbook of Market Research*.
- Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., & Ray, S. (2019). Prediction-oriented model selection in partial least squares path modeling. *Decision Sciences*. <https://doi.org/10.1111/deci.12329>. (forthcoming).
- Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research*, 69(10), 4552–4564.
- Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J.-H., Ting, H., Vaithilingam, S., et al. (2019). Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *European Journal of Marketing*, forthcoming.
- Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational statistics and data analysis*, 159–205.
- Tenenhaus, A., & Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2), 257–284.
- Trincherà, L., Marie, N., & Marcoulides, G. A. (2018). A distribution free interval estimate for Coefficient Alpha. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(6), 876–887.
- Wang, Y., Feng, X. N., & Song, X. Y. (2016). Bayesian quantile structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2), 246–258.
- Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. Jöreskog & H. Wold (Eds.), *Systems under indirect observation* (Vol. 2, pp. 1–54). Amsterdam: North-Holland.
- Wold, H. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*. Hoboken: Wiley.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.