Springer Proceedings in Mathematics & Statistics

Nicola Salvati
Cira Perna
Stefano Marchetti
Raymond Chambers *Editors*

Studies in Theoretical and Applied Statistics

SIS 2021, Pisa, Italy, June 21–25

SIS Società Italiana di Statistica

Springer

Springer Proceedings in Mathematics & Statistics

Volume 406

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Nicola Salvati · Cira Perna · Stefano Marchetti · Raymond Chambers
Editors

Studies in Theoretical and Applied Statistics

SIS 2021, Pisa, Italy, June 21–25

Springer

Editors
Nicola Salvati
Department of Economics and Management
University of Pisa
Pisa, Italy

Cira Perna
Department of Economics and Statistics
University of Salerno
Fisciano, Salerno, Italy

Stefano Marchetti
Department of Economics and Management
University of Pisa
Pisa, Italy

Raymond Chambers
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong, NSW, Australia

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-031-16608-2 ISBN 978-3-031-16609-9 (eBook)
https://doi.org/10.1007/978-3-031-16609-9

Mathematics Subject Classification: 62-06

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book gathers selected peer-reviewed papers presented during the 50th Scientific Meeting of the Italian Statistical Society (SIS2021). Due to the Covid-19 pandemic, which limited the mobility of the staff of many universities and research centres, SIS2021 was conducted remotely from the 21st to the 25th of June 2021.

This biennial conference is a traditional meeting for promoting interactions among national and international researchers in statistics, demography, and applied statistics in Italy. The aim of the conference is to bring together national and foreign researchers and practitioners to discuss recent developments in theoretical and applied statistics as well as in demography and statistics for the social sciences.

The Scientific Program Committee and the Organizing Committee of SIS2021 put together a balanced and stimulating program which was of great interest to all participants.

The conference program included 4 plenary sessions, 15 specialized sessions, 20 solicited sessions, 37 contributed sessions, and the poster exhibition. The meeting also hosted three Satellite Events on 'Measuring uncertainty in key official economic statistics', 'Covid-19: the urgent call for a unified statistical and demographic challenge', and 'Evento SIS-PLS Statistica in classe: verso un insegnamento laboratoriale'. There were 323 submissions accepted by the Scientific Program Committee, including 128 that were presented at invited plenary, specialized and solicited sessions, and 195 that were submitted as contributed papers for oral presentation and for the poster sessions.

This book of selected papers from those presented at SIS2021 covers a wide variety of subjects and provides an overview of the current state of Italian scientific research in theoretical and applied statistics. The papers contained in this book cover areas that include Bayesian models, survey methods, time series models, spatial models, finance models, clustering methods, and new methods and applications to Covid-19.

The Scientific Program Committee, the Organizing Committee, and many volunteers contributed to the organization of SIS2021 and to the refereeing of the papers included in this book. Our heartfelt thanks go to all of them. A special thank you

goes to Francesco Schirripa Spagnolo for his continuous assistance and support in the organization of the conference and in the editing of this book.

Wishing you a productive and stimulating reading experience.

Pisa, Italy — Nicola Salvati
Salerno, Italy — Cira Perna
Pisa, Italy — Stefano Marchetti
Wollongong, Australia — Raymond Chambers

Contents

A Composite Index of Economic Well-Being for the European Union Countries

Andrea Cutillo, Matteo Mazziotta, and Adriano Pareto

Abstract The measurement of Equitable and Sustainable Well-being (BES) in Italy is one of the most appreciated monitoring tools by the Scientific Community. The focus on the Economic Well-being domain seems essential around the last serious economic crisis. The use of an innovative composite index can help to measure the multidimensional phenomenon and monitor the situation at European level.

Keywords Composite index · Ranking · Economic well-being

1 Introduction

In this paper, the economic well-being in Europe is focused, taking as a reference point the economic domain of the project BES (Equitable and Sustainable Well-being in Italy) of the Italian National Institute of Statistics (Istat). The BES aims at evaluating the progress of societies by considering different perspectives through twelve relevant theoretical domains, each one measured through a different set of individual indicators. The BES project is inspired by the Global Project on Measuring the progress of Societies of the Oecd (2007), with the idea that the economic well-being is not enough for the developed Countries. However, since 2007, two huge economic crises have affected the households' economic well-being: the international economic crisis (about 2008–2009) derived from the Lehman Brothers failure; and

A. Cutillo · M. Mazziotta (✉) · A. Pareto
Italian National Institute of Statistics, Rome, Italy
e-mail: mazziott@istat.it

A. Cutillo
e-mail: cutillo@istat.it

A. Pareto
e-mail: pareto@istat.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_1

the European crisis of the sovereign debts, whose effects were more intense in 2011–2012, and can be considered solved in 2014.[1] In the meantime, the EU fiscal and monetary policies have completely changed, going from very restrictive ones in the international economic crisis and in the first part of the sovereign debt crisis, to more expansive ones (especially the monetary policy) starting from the second part of the sovereign debt crisis till nowadays. This fact has reflected in a great improve of the European household' economic conditions, as can be seen in the next paragraphs. Then, the economic domain of the well-being still deserves particular relevance within the other dimensions. Following the timeliness described above, the longitudinal analysis is set at 4 relevant years: 2007, 2010, 2014 and 2019.

2 Theoretical Framework

The starting point of our framework is the Istat BES: it measures the economic domain through a set of ten indicators. However, we operate some changes due to operational issues (data availability and comparability with the other countries for the wealth indicators and the absolute poverty indicator) as well as theoretical issues (a couple of indicators can hardly be considered as economic well-being indicators and another one has been excluded in order to avoid a double counting of inequality). Changes and restriction are extensively described in the depiction of the adopted sub-domains and indicators. In this paper, we measure the economic well-being through four sub-domains (Purchasing power, Inequality, Poverty and Subjective evaluation), each one represented by a single indicator coming from the Eu-Silc (European Statistics on Income and Living Conditions) system. Purchasing power can give an evaluation of the average economic standard of a Country; Inequality is an important issue even in case of the rich Countries, since it measures the share of people who are relatively disadvantaged in respect of their social and economic context; poverty measures the share of people who can't reach a minimum standard of living; and the subjective evaluation is important in order to capture people who feel to have economic problems, even when they do not have difficulties under an objective point of view.

1. Sub-domain *Purchasing Power*; indicator: *Median equivalised income in purchasing power standards (Pps).* The Istat *average income per capita* is replaced for three reasons. First, the median is a better indicator of a monetary distribution, given its robustness to extreme values. Second, the equivalised form (through the modified Oecd scale) is better in order to consider the different sizes and needs of the households. In the opinion of the authors, also the Istat BES could benefit in changing accordingly. Finally, in the European context, it is essential to consider the different cost of life and purchasing powers in the

[1] Obviously, we cannot forget the current crisis deriving from the Covid19 pandemic situation. However, the adopted indicators are not still available for 2020 in all the Countries. Moreover, it should be evaluated when the pandemic situation will be officially declared as finished.

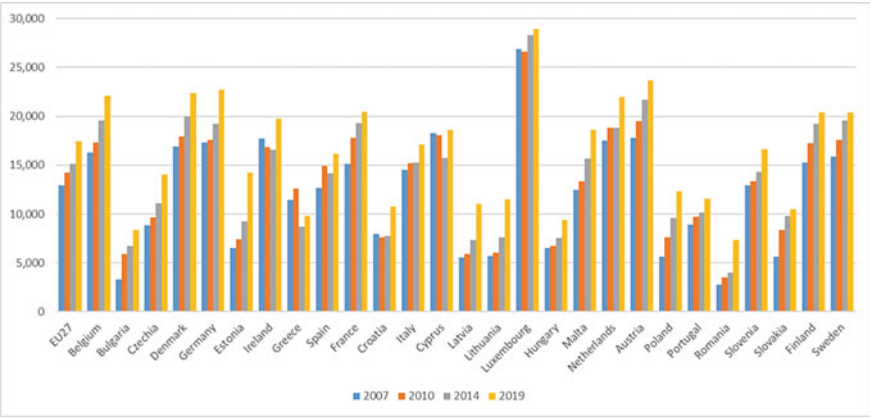


Fig. 1 Median equivalised income in Pps in the EU27 Countries. Years 2007, 2010, 2014 and 2019. Values in euros

Countries. The values of this indicator (Fig. 1) positively defined in respect of well-being, range between 2,783 euros (Romania in 2007) to 28,943 (Luxemburg in 2019). Romania presents the lowest values in all the years, even if this Country multiplies its value by 2.6 in the entire period (7,338 euros in 2019), while Luxemburg presents the highest values in all the years. The values in the entire EU27 are 12,927 euros in 2007, 14,235 in 2010, 15,137 in 2014 and 17,422 euros in 2019.

2. Sub-domain *Inequality*; indicator: *At risk of poverty rate (ARP).* It is a relative measure of poverty: its threshold is set dependently on the income distribution and, therefore, it merely captures how many individuals are far from the others. That is, relative poverty is an inequality indicator rather than a poverty indicator [7]. Istat measures inequality also through the *Disposable income inequality* (S80_S20 index, which is the ratio of total equivalised income received by the 20% of the population with the highest income to that received by the 20% of the population with the lowest income). Since they are both representative of the same sub-domain and in order to avoid a double counting of the same domain, only the ARP is selected (as a matter of fact, the correlation coefficient between the two indicators is more than 0.90). As a general rule, it is a good practice to strictly select indicators in the construction of composite indicators. The ARP generally shows the lowest degree of variability across the Countries as well as across the years (Fig. 2). The values of ARP, negatively defined, range between 9.0% (Czechia in 2010) to 25.1% (Romania in 2014). Czechia presents the lowest values in all the years, while Romania presents the highest values in all the years. The values in the entire EU27 are 16.3% in 2007, 16.5% in 2010, 17.3% in 2014 and 16.5% euros in 2019.
3. Sub-domain *Poverty*; indicator *Severe material deprivation (SMD)*, that is the share of population living in households lacking at least 4 items out of 9 economic

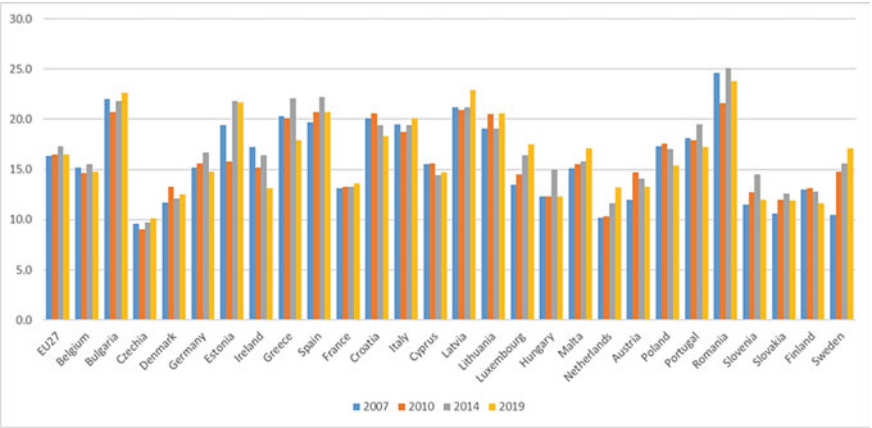


Fig. 2 At risk of poverty rate in the EU27 Countries. Years 2007, 2010, 2014 and 2019. Values in percentages

deprivations. Far from being a perfect indicator, it is the most similar indicator to the concept of absolute poverty in the EU. Unwillingly, Istat *Absolute poverty rate* cannot be used even if it is a better measure of poverty (the poverty lines are set independently of the monetary distribution, and also consider the different cost of life in different areas). However, the absolute poverty is officially measured only in Italy and USA, because of the difficulties in its definition, and the European Commission project "Measuring and monitoring absolute poverty—ABSPO" is still in the phase of study [4]. Leaving aside the World Bank measure, which does not fit for developed Countries, the SMD is the only indicator that permits European comparison in this domain, entailing data comparability. The values of this indicator (Fig. 3), negatively defined, range between 0.5% (Luxemburg in 2010) to 57.6% (Bulgaria in 2007). Bulgaria presents the highest values in all the years, but also shows a dramatic fall in the course of the years (19.9% in 2019), partly filling the gap with the other Countries. The values in the entire EU27 are 9.8% in 2007, 8.9% in 2010, 9.1% in 2014 and 5.4% in 2019.

4. Sub-domain *Subjective evaluation*; the indicator *Index of self-reported economic distress*, that is the share of individuals who declare to get to the end of the month with great difficulty. The subjective sub-dimension is considered an important one, since it can capture a worsening in well-being for people who feel to have economic problems, even when they have not difficulties with an objective point of view. This is particularly relevant especially in years in which the economic crises could have highly changed the perceptions of the households in a different way between Countries. The values of this indicator (Fig. 4), negatively defined, range between 1.4% (Germany in 2019) to 39.5% (Greece in 2014). Indeed, Germany is the Country that appeared as the leading one in EU in these years, and this fact reflects in the perceptions of the households. At the opposite, the dramatic jump in the 2014 Greek indicator indicates the uncertainty deriving

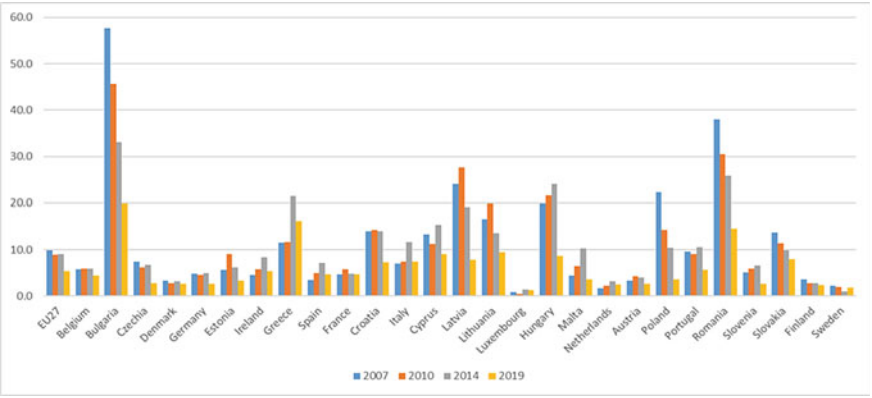


Fig. 3 Severe material deprivation in the EU27 Countries. Years 2007, 2010, 2014 and 2019. Values in percentage

from the consequences of the crisis for the Greek households. The values in the entire EU27 are 9.8% in 2007, 11.2% in 2010, 11.8% in 2014 and 6.5% in 2019.

The remaining four Istat indicators are removed for the following reasons: *Per capita net wealth*: the sub-domain wealth is certainly a pillar of the households' monetary well-being. However, correctly measuring the value of wealth is extremely complex [1], since some types of wealth are statistically hidden (e.g., paintings, jewellery etc.), and attributing a value to wealth is arbitrary when some types of wealth, e.g. houses, are not sold/bought. Unfortunately, this exclusion is a relevant issue in the European context, considering the different weight between financial

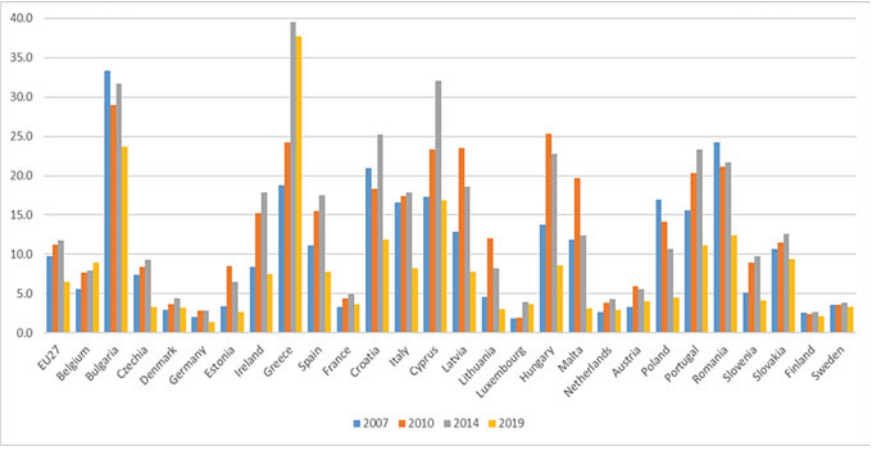


Fig. 4 Households declaring to get to the end of the month with great difficulty in the EU27 Countries. Years 2007, 2010, 2014 and 2019. Values in percentage

wealth and real estate wealth in the different countries. *People living in financially vulnerable households*, measured through the percentage of households with debt service greater than 30% of disposable income: to the best of our knowledge, there is not such indicator in the Eu-Silc database. *Severe housing deprivation* (Share of population living in overcrowded dwellings and also exhibits at least one of some structural problem) and *Low work intensity* (Proportion of people 0–59 living in households in which household members of working age worked less than 20% of the number of months that could theoretically have been worked) measure important topics, but, according to our views, they can't be considered as indicators of economic well-being from a theoretical point of view.

3 Methodological Aspects

The composite index was constructed using the Adjusted Mazziotta-Pareto Index—AMPI [5]. This aggregation function allows a partial compensability, so that an increase in the most deprived indicator will have a higher impact on the composite index (imperfect substitutability). Such a choice is advisable whenever a reasonable achievement in any of the individual indicators is considered to be crucial for overall performance [3]. The most original aspect of this index is the method of normalization, called "Constrained Min–Max Method" [6]. This method normalizes the range of individual indicators, similarly to the classic Min–Max method, but uses a common reference that allows to define a 'balancing model' (i.e., the set of values that are considered balanced). Thus, it is possible to compare the values of the units, both in space and time, with respect to a common reference that does not change over time.

Let us consider the matrix $\mathbf{X} = \{x_{ijt}\}$ with 27 rows (countries), 4 columns (individual indicators), and 4 layers (years) where x_{ijt} is the value of individual indicator j, for country i, at year t. A normalized matrix $\mathbf{R} = \{r_{ijt}\}$ is computed as follows:

$$r_{ijt} = 100 \pm \frac{x_{ijt} - x_{j0}}{\max\limits_{it}(x_{ijt}) - \min\limits_{it}(x_{ijt})} 60$$

where $\min\limits_{it}(x_{ijt})$ and $\max\limits_{it}(x_{ijt})$ are, respectively, the overall minimum and maximum of indicator j across all times (goalposts), x_{j0} is the EU average in 2007 (reference value) for indicator j, and the sign $\pm$ depends on the polarity of indicator j.

Denoting with $\mathrm{M}_{r_{it}}$, $\mathrm{S}_{r_{it}}$, $\mathrm{cv}_{r_{it}}$, respectively, the mean, standard deviation, and coefficient of variation of the normalized values for country i, at year t, the composite index is given by:

$$\mathrm{AMPI}_{it}^{-} = \mathrm{M}_{r_{it}} - \mathrm{S}_{r_{it}} \mathrm{cv}_{r_{it}}$$

where:

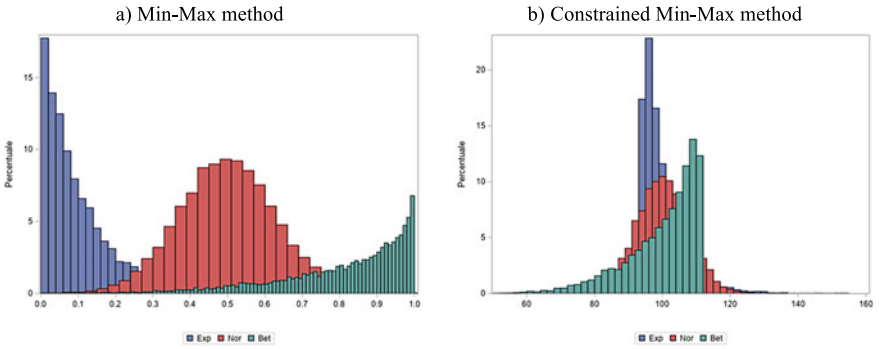


Fig. 5 Comparing the classic and the constrained Min–Max method

$$\mathrm{M}_{r_{it}} = \frac{\sum_{j=1}^{4} r_{ijt}}{4} \quad \mathrm{S}_{r_{it}} = \sqrt{\frac{\sum_{j=1}^{4} (r_{ijt} - \mathrm{M}_{r_{it}})^2}{4}} \quad \mathrm{cv}_{r_{it}} = \frac{\mathrm{S}_{r_{it}}}{\mathrm{M}_{r_{it}}}.$$

The version of AMPI with a negative penalty was used, as the composite index is 'increasing' or 'positive', i.e., increasing values of the index correspond to positive variations of the economic well-being. Therefore, an unbalance among indicators will have a negative effect on the value of the index [5].

Figure 5 shows the effect of normalization on three individual indicators with different shape generated in a simulation.[2] The first has an exponential distribution with $\lambda = 0.0125$ (Exp), the second has a normal distribution with $\mu = 150$ and $\sigma = 15$ (Nor) and the third has a Beta distribution with $\alpha = 4$ and $\beta = 0.8$ (Bet). The indicators have different parameters, as they represent the most disparate phenomena. In Fig. 5a, indicators are normalized by the classic Min–Max method in the range [0, 1], and in Fig. 5b, they are normalized by the constrained Min–max method with a reference (the mean) of 100 and a range of 60.

As we can see, the Min–Max method bring all values into a closed interval, but the distributions of indicators are not 'centred' and this leads to the loss of a common reference value, such as the mean. It follows that equal normalized values (i.e., balanced normalized values) can correspond to very unbalanced original values. For example, the normalized value 0.2 for the Exp indicator corresponds to a high original value; whereas for the Nor and Bet indicators it corresponds to a very low original value. Moreover, the normalized value 0.5 is the mean of the range, but not of distributions, and then it cannot be used as a reference for reading results (e.g., if the normalized value of a country is 0.3., we cannot know if its original value is

[2] Note that socio-economic indicators are basically of two types: per capita indicators and percentage indicators. Per capita indicators tend to be open-ended, in particular at the upper end of the range (e.g., *GDP per capita*); percentage type indicators tend to have severe constraints operating at the upper end of the range, with consequent piling up of observation there (e.g., *Adult literacy*). Therefore, most of individual indicators have positively or negatively skewed distributions [2].

Fig. 6 AMPI indicator. Distance from the reference value (EU27 in 2007 = 100) in the EU27 countries. Years 2007, 2010, 2014 and 2019

above or below the mean). On the other hand, normalized values by the constrained Min–Max method are not forced into a closed interval, they are 'centred' with respect to a common reference, and they are easier to interpret: if the normalized value of a country is greater than 100, then it is above the reference value, else it is below the reference value (Fig. 6). Finally, the comparability across time is maintained when new data become available (the goalposts do not need to be updated).

4 A Longitudinal Analysis

In the analysis, the reference value is the Eu27 in 2007 (=100), and each value can be evaluated as the relative distance to 100 (Table 1). The Eu27 indicator is not far from 100 neither in 2010 (100.2) nor in 2014 (99.6). The last year shows instead an increase of the overall index of about 5 point (104.7 in 2019).

Before commenting the different phases, it can be of interest to observe which indicators have the greatest impact in the AMPI. All the four primary indicators (one positively defined and three negatively defined in respect of economic wellbeing) are obviously highly correlated with the AMPI. However, the one that shows the highest correlation is the poverty indicator (severe material deprivation), about −0.90 in the four years, while the one with the lowest correlation is the inequality index (at risk of poverty rate), that decreases from −0.80 in 2007 to −0.75 in 2019 (Table 2).

The first phase, corresponding to the international economic crisis, is the most stable. Indeed, the ranking, based on the AMPI, shows a low level of variability between 2007 and 2010, as well as the values of the AMPI. The highest jump in the AMPI absolute value is observed for Poland, which also passes from the 23rd

Table 1 AMPI value and ranking in the EU Countries, years 2007, 2010, 2014 and 2017

Country	Year							
	2007		2010		2014		2019	
	AMPI	Rank	AMPI	Rank	AMPI	Rank	AMPI	Rank
Belgium	105.7	10	105.9	9	106.1	9	107.9	11
Bulgaria	65.3	27	73.7	27	75.9	26	83.0	26
Czechia	104.0	11	104.9	10	105.0	10	110.3	8
Denmark	110.9	5	109.9	4	111.7	2	113.2	2
Germany	107.8	8	107.4	8	106.9	8	111.8	5
Estonia	95.9	17	97.7	14	93.9	17	98.8	21
Ireland	103.6	12	101.8	12	98.6	14	108.8	10
Greece	91.1	22	89.2	21	74.6	27	80.9	27
Spain	97.4	16	95.3	18	91.7	19	99.0	19
France	108.1	7	108.7	5	109.6	6	110.4	7
Croatia	87.9	24	88.2	24	86.4	23	96.6	23
Italy	95.4	18	96.2	17	94.2	16	99.4	18
Cyprus	99.2	14	96.4	16	90.0	20	101.4	17
Latvia	86.0	25	81.6	25	86.2	24	92.8	24
Lithuania	92.7	21	88.3	23	93.6	18	97.1	22
Luxembourg	115.4	1	114.3	1	111.7	1	110.7	6
Hungary	94.5	19	88.6	22	88.3	22	101.8	16
Malta	101.3	13	97.4	15	100.8	12	106.4	13
Netherlands	113.0	2	113.1	2	111.5	3	112.5	4
Austria	111.0	4	108.1	7	110.0	5	112.8	3
Poland	88.5	23	92.8	19	96.8	15	104.1	14
Portugal	93.7	20	92.4	20	89.3	21	98.9	20
Romania	73.2	26	79.8	26	77.3	25	86.5	25
Slovenia	107.2	9	104.8	11	103.3	11	110.2	9
Slovakia	97.9	15	99.4	13	99.9	13	102.8	15
Finland	108.8	6	110.2	3	111.5	4	113.5	1
Sweden	111.3	3	108.6	6	109.0	7	107.7	12
EU27	**100.0**		**100.2**		**99.6**		**104.7**	

position to the 19th in 2010. In the second phase, corresponding to the crisis of the sovereign debt, there is a greater mobility in the ranking. Greece shows the highest jump, from 21st to 27th and last position. The Greek AMPI decreased dramatically from 89.2 to 74.6. This fall was mainly due to a dramatic fall in the purchasing power of the households (the median equivalised income in Pps decreased from 12,598 to 8,673 euros). Also, the SMD and the subjective economic distress greatly worsened, respectively from 11.6% to 21.5% and from 24.2% to 39.5%). Indeed Greece was

Table 2 Correlation coefficient between the primary indicators and the AMPI in the different years

Indicator	Year			
	2007	2010	2014	2019
Median equivalised income in pps	0.82	0.83	0.83	0.80
At risk of poverty rate	−0.80	−0.76	−0.79	−0.75
Severe material deprivation	−0.91	−0.90	−0.90	−0.90
Index of self-reported economic distress	−0.90	−0.89	−0.89	−0.79

the first country to be hit by the equity markets distrust on the debt sustainability, later followed by Portugal and Ireland and successively by Italy and Spain. In the 2010–2014 phase, Ireland loses two positions (from 12 to 14th), Spain and Portugal one position (respectively, from 18 to 19th and from 21st to 22nd), while Italy gained one position (from 17 to 16th). However, also Italy showed a decrease in the synthetic index, from 96.2 to 94.2 and the overall Italian situation was somewhat preserved by the fact that only the SMD indicator worsened (from 7.4 to 11.6%), while the other three were substantially unchanged. In this time, we can observe a new great advance of Poland (+4 in the ranking, from 19 to 15th), which shows an increase of the MPI from 92.8 to 96.8.

In the opinion of the authors, these data clearly show that the European response to the sovereign debt crisis has done more harm than good. The vexatious conditions imposed to Greece by the European Commission, European Central Bank and International Monetary Fund highly worsened the household economic conditions of the Country and were badly used as a warning for the other indebted Countries. Unsurprisingly, they were instead used by the stock markets' operators as a sign of permit towards speculation, which quickly enlarged against the other Countries. Luckily, when the entire Eurozone was in doubt, the European institutions changed their policies. IMF was involved less intensely; the ECB completely changed its monetary policy, which originally just looked at an about non-existent inflation and did not foresee an intervention on the stock markets (the Quantitative Easing started in 2012 in order to support the financial system and to save the Euro area; somewhat enlarged its effects on the productivity system in 2014; and started its second and stronger phase in 2015, with an always greater intervention); and the Eurozone, even in a context of a formally stricter balance observation through the *fiscal compact*, contemplated a series of adjustments which allowed to keep in account a number of factors (e.g., the years of general economic crisis as well as the notion of "potential GDP") rather than applying in aseptic way the treaties. The new policies facilitated the growth of the GDP as well as an improvement in the households' economic conditions in Europe, as observed in the data. Indeed, in the last phase, till 2019, the overall Eu27 index passed from 99.6 to 104.7, showing a general increase on the economic well-being of the households, and all the Countries, but Sweden and Luxemburg, increased the value of the index. Some Countries had a particularly great increase (Hungary, Cyprus, Croatia and Ireland, more than +10 points). As concerning the ranking, Hungary showed the greatest increase, + 6 positions, especially due to an

improvement in median purchasing power, SMD and subjective economic distress; Luxemburg and Sweden showed the greatest decrease, −5 positions, especially due to an increase of inequality as measured by the ARP rate in a context of general decrease of inequality in the European zone.

Considering the entire time frame, 2007–2019, some Countries greatly increased their economic well-being, in particular, and somewhat obviously given that the economic convergence is one of the targets of the EU, the Countries that started from a disadvantaged situation: Bulgaria (+17.7), Poland (+15.6) and Romania (+13.3). In the case of Poland, this also pushed the ranking, from the 23rd to the 14th position; Bulgaria and Romania still remain at the bottom tail of the ranking, respectively 26th and 25th in 2019, +1 position for both the Countries, but strongly filled the gap in respect of the overall EU27. At the opposite, the Greek indicator has fallen down by 10.1 point (even if it is growing in the last sub-time), completely due to the 2011–2014 time frame. At present, Greece is in the last position of the ranking, 27th (vs 22nd in 2007), while the first position is occupied by Finland. The other two Countries with an important decrease in the MPI indicator are Luxemburg, −4.8 points, and Sweden, −3.5 points, which shifted, respectively, from 1st to 6th position and from 3rd to 12th position.

Summarizing, the overall time frame is divided in the following phases of the international economic crisis: 2007–2010; the phase of the Eurozone crisis, 2011–2014; and the last phase of economic stability, 2015–2019. The first two appear to be a long period of unique crisis, slightly softer and more diffuse in the first part; more intense and localized in a fewer number of Countries in the second part. The particularity of this second phase is that even the Countries which didn't face the crisis of the sovereign debts didn't improve their economic well-being, showing that the entire Europe has faced it as well, and the only way for advancing is solidarity and reasonability. The last phase was indeed characterized by a general increase of the European households' economic well-being, mostly due to a more reasonable and rational use of the fiscal policies and, especially, of the monetary policies. Such measures have permitted to relax the economic distress on the European households.

As concerning Italy, it started 18th in 2007 and is 18th in 2019, with a negative gap in respect of the EU27 which ranged from −4 to −5.5 points in the phase. Looking at the sub-domains, Italy has improved its purchasing power, even though with a grow rate lower than the Eu27; the ARP is very stable along the phase; the SMD indicator shows a high value only in 2014 (11.6% vs a little more than 7% in the other years); the subjective economic distress about halves in the last phase (from 17.9% to 8.2%), even due to a more stable economic situation which reflects on the opinion of the households.

5 Conclusions

In this paper we analysed the economic well-being in Europe in the 2007–2019 time frame at 4 relevant years: 2007, 2010, 2014 and 2019. In order to do so, we have considered the economic domain of the project BES (Equitable and Sustainable Well-being) of the Italian National Institute of Statistics (Istat), changing its definition in accordance with data comparability and theoretical issues. Such domain was measured through four sub-domains: *Purchasing Power is measured by* the *Median equivalised income in purchasing power standards (Pps). Inequality is measured by* the *At risk of poverty rate (ARP). Poverty* is measured by the *Severe material deprivation (SMD);* and *Subjective evaluation* is measured by the *Index of economic distress.*

Following the aforementioned years, the analyses considers 3 relevant phases: 2007–2010, which comprises the international economic crisis; 2011–2014, which comprises the crisis of the European sovereign debts; and 2015–2019, that is characterized by relative stability and recovery. The first two phases appear to be a unique long time frame of crisis, slightly softer and more diffuse in the first part; more intense and localized in a fewer number of Countries in the second part, particularly heavy for Greece. The peculiarity of this second phase is that even the Countries which didn't face the crisis of the sovereign debts didn't improve their economic well-being. This fact clearly show that the European response was far from being satisfactory, and the vexatious conditions imposed to Greece by EC, ECB and IMF highly worsened the Greek household economic conditions and were badly used as a warning for other indebted Countries. On the other hand, such measures stimulated the stock markets' speculation, which quickly enlarged against the other Countries, and the entire Eurozone was put in doubt. Luckily, fiscal and monetary policies have completely changed since then. The IMF was involved less intensely; the Eurozone, even in a context of a formally stricter balance observation through the fiscal compact, contemplated several adjustments which allowed to keep in account different factors; and, mainly, the ECB completely changed its monetary policy (through the quantitative easing that started in 2012 and enlarged its dimension starting from 2015). Such measures have permitted to relax the economic distress on the European households and all European countries have resumed the normal path towards the higher and generalized economic well-being that characterized the whole post-war period. In this regard, it has to be noted that Germany, the Country that undoubtedly has leaded the EU in the considered time frame, does not improve its position neither in the value of the indicator nor in the ranking (8th) till 2014, while it increased the value of the indicator (+4.9 points) and the ranking (+3 positions) in the last phase, when "less German" fiscal and monetary policies were applied (certainly with the agreement of Germany itself), as a further confirmation of the fact that the only way for economic advancing in EU is solidarity and reasonability.

References

1. Canberra Group: Handbook on Household Income Statistics. United Nations, Geneva (2011)
2. Casacci, S., Pareto, A.: A nonlinear multivariate method for constructing profiles of country performance. Int. J. Multicriteria Decis. Mak. **8**, 295–312 (2021)
3. Chiappero-Martinetti, E., von Jacobi, N.: Light and shade of multidimensional indexes. How methodological choices impact on empirical results. In: F. Maggino, G. Nuvolati (eds), Quality of life in Italy, Research and Reflections, pp. 69–103, Springer, Cham (2012)
4. Cutillo, A., Raitano, M., Siciliani, I.: Income-based and consumption-based measurement of absolute poverty: insights from Italy. Soc. Indic. Res. (2020). https://doi.org/10.1007/s11205-020-02386-9
5. Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. Soc. Indic. Res. **127**, 983–1003 (2016)
6. Mazziotta, M., Pareto, A.: Everything you always wanted to know about normalization (but were afraid to ask). Ital. Rev. Econ., Demogr. Stat., LXXV **1**, 41–52 (2021)
7. Sen, A.: Poor, relatively speaking. Oxford Econ. Ser. **35**(2), 153–169 (1983)

A Dynamic Power Prior for Bayesian Non-inferiority Trials

Fulvio De Santis and Stefania Gubbiotti

Abstract Non-inferiority trials compare new experimental treatments to active controls. Previous information on the control treatments is often available and, as long as the past and the current experiments are sufficiently homogeneous, historical data may be useful to reserve resources to the new therapy's arm and to improve accuracy of inference. In this article we propose a Bayesian method for exploiting historical information based on a dynamic power prior for the parameter of the control arm. The degree of information-borrowing is tuned by a quantity based on the Hellinger distance between the two posterior distributions of the control arm's parameter, obtained respectively from the current and the historical experiments. Pre-posterior analysis for type-I error/power assessment and for sample size determination is also discussed.

Keywords Clinical trials · Hellinger distance · Historical data · Power prior · Sample size determination

1 Introduction

A Non-inferiority (NI) trial is an experiment where a new treatment is compared to an existing active therapy (control). Unlike trials in which a new effective treatment must be shown to be superior to the placebo, the objective of a NI trial is to establish that the difference between the effects of the new and the control treatments is small enough to conclude that the new drug is also effective. NI trials are therefore typically used to draw inference on the unknown parameter: if the hypothesis that the unknown parameter is below a given *non-inferiority margin* is rejected, one can conclude for NI. The importance of "borrowing" information from previous studies is that "with

F. De Santis · S. Gubbiotti (✉)
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, Piazzale Aldo Moro N. 5, 00185 Roma, Italy
e-mail: stefania.gubbiotti@uniroma1.it

F. De Santis
e-mail: fulvio.desantis@uniroma1.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_2

historical data providing information on the control arm, more trial resources can be devoted to the novel treatment while retaining accurate estimates of the current control arm parameters" [28]. This may result in more accurate inference, as long as historical information is sufficiently similar to the current control data. Techniques to borrow information from historical data have been developed both from the frequentist and the Bayesian perspectives. One advantage of the Bayesian approach is that it allows a very natural way to exploit this historical information on the control parameter: data from previous experiments can in fact be used to define a distribution for the effect of the control treatment to be employed as prior in the current experiment. In this regard, see, among others [4, 8–11, 18, 19, 22]. Bayesian methodology also provides several approaches to discount the level of borrowing from historical data. These methods take into account the degree of compatibility between data from current and past experiments. Such a problem has been addressed, for instance, by [14–16, 20, 28]. Among the available alternative borrowing approaches, in this article we focus on the power priors methodology. The idea, originally proposed in the seminal paper [17], prescribes to define the prior for the control parameter to be proportional to a starting density (typically a non-informative prior) times the likelihood associated to historical data raised to a parameter that ranges in [0, 1] and weights the historical data relative to the likelihood of the current study. The power prior parameter tunes the influence of the past data on the distribution of the control parameter: a value equal to 0 is equivalent to no incorporation of historical data in the prior; a value equal to 1 corresponds to full borrowing; intermediate values imply partial borrowing. The choice is then crucial especially when there is heterogeneity between the previous and the current trial or when the sample sizes of the two studies are significantly different. Several methods have been proposed to deal with this choice. In addition to the two main strategies, that consider either a fixed value or a random variable with a given density on the unit interval, some authors have recently proposed to consider a function of a measure of congruence between historical and current data. This approach yields the so-called dynamic power prior: see, for instance, [12, 13, 19, 21, 23, 24].

The present article is a wholly Bayesian conversion of the hybrid frequentist-Bayes method proposed by [19] borrowing ideas from [23]. The main features of Liu's approach in [19] are: (i) implementation of a frequentist test for NI; (ii) instrumental use of a dynamic power prior *only* for the selection of the amount of borrowing from historical data (no posterior analysis is considered); (iii) definition of the power prior parameter as an arbitrary function of the p-value for testing the hypothesis of equivalence between the current and historical control true response rates. Features (ii) and (iii) present some controversial aspects. Specifically, for (ii) one can object that an instrumental use of the power prior does not have a clear justification outside a Bayesian context; with respect to (iii), one can call into question the arbitrary choice of the p-value function that may yield any value of the power prior parameter in [0, 1]. For these reasons, in this paper we propose: (i) to make use of a Bayesian test of NI, based on a credible interval for the unknown effects difference; (ii) to consider a power prior to build the posterior distributions of the parameter necessary for feature (i); (iii) to define a new dynamic fraction based on a sensible measure of

compatibility between historical and current data using the Hellinger distance (see [23]).

The outline of the paper is as follows. In Sect. 2 we describe the Bayesian methodology that we propose for the NI test. Section 2.1 provides details on the construction of the priors and on the derivation of the corresponding posteriors. Specifically, we assume the dynamic fraction to be a function of the Hellinger distance between the posterior densities of the control parameter given the current and historical data of the control arms, respectively. We also consider the possibility of setting an upper bound to the amount of information borrowed from previous studies in order to avoid that current data is overwhelmed by pre-experimental information. We explore in Sect. 3 the main (posterior) features of the proposed approach in a real NI study on vaccine considered in [19]. In order to address the requirements of regulatory agencies [2, 7], in Sect. 4 we also investigate frequentist properties of our proposal in terms of type-I error and power. In Sect. 5 we introduce a sample size determination criterion. Discussions on Bayesian experimental design and sample size determination can be found, among others, in [1, 4–6, 10, 11, 18, 25, 27, 29]. Finally Sect. 6 contains a discussion.

2 Methodology

Let us consider a two-arms trial where an experimental drug (e) is compared to a standard therapy, here used as control (c). Let θ_e and θ_c denote the corresponding unknown probabilities of success and let X_e and X_c denote the random number of positive responses out of n_e and n_c observations in the two arms. We assume that $X_j|\theta_j \sim Bin(n_j, \theta_j)$, $j = e, c$ and that $X_e \perp X_c \,|\, \theta_e, \theta_c$. Non-inferiority of drug e with respect to drug c is assessed if the null hypothesis of the test

$$H_0 : \theta_e - \theta_c \leq -\delta \quad vs. \quad H_1 : \theta_e - \theta_c > -\delta \tag{1}$$

is rejected, where $\delta > 0$ is a selected NI margin. Adopting the Bayesian paradigm, we proceed as follows. We determine a credible interval $C = [L, U]$ for $\theta = \theta_e - \theta_c$ of level $1 - \gamma$ and we reject H_0 if $L > -\delta$. Determination of C requires the posterior distributions of θ_e and θ_c. We assume that no information on θ_e is available, whereas historical data regarding θ_c can be retrieved. Under these assumptions we construct the prior distributions for θ_e and θ_c and derive the corresponding posteriors, as detailed in the following subsection. Based on these posterior distributions, the lower limit L of the equal tails interval for $\theta = \theta_e - \theta_c$ is simply computed via Monte Carlo. Then, if $L > -\delta$ the null hypotesis of the NI test is rejected.

2.1 Priors Construction

Let $\pi_e(\cdot)$ be the non-informative $Beta(1, 1)$ density prior for θ_e. Given the experimental data x_e and using $\pi_e(\cdot)$ we obtain the posterior $\pi_e(\cdot|x_e)$ that is a $Beta(1 + x_e, 1 + n_e - x_e)$ density. Furthermore, let us assume that a previous study provides historical data (n_h, x_h) yielding information on the control parameter θ_c, where n_h and x_h are the size and the number of successes. As prior for θ_c in the current experiment we then consider its posterior density given x_h. In order to take into account potential heterogeneity between current and historical information on θ_c, we consider the power prior originally defined by [17] as

$$\pi_c^P(\theta_c|x_h) \propto \pi_c^o(\theta_c) \times [f(x_h|\theta_c)]^a, \quad a \in [0, 1] \tag{2}$$

where $\pi_c^o(\theta_c)$ is a starting prior (typically a non-informative prior), $f(x_h|\theta_c)$ the likelihood function of θ_c given the historical data x_h and $a \in [0, 1]$ a discount parameter. The smaller a, the lighter the degree of incorporation of historical information: $a = 0$ corresponds to no borrowing, whereas $a = 1$ implies full borrowing. Noting that $[f(x_h|\theta_c)]^a \propto \theta_c^{ax_h}(1 - \theta_c)^{a(n_h - x_h)}$ and assuming $\pi_c^o(\cdot)$ to be the $Beta(1, 1)$ density, we have that $\pi_c^P(\theta_c|x_h, x_c)$ is the $Beta(1 + ax_h + x_c, 1 + a(n_h - x_h) + n_c - x_c)$ density.

The choice of a is crucial in determining the impact of historical data on the analysis. As an extreme case, if x_h and x_c can be considered fully exchangeable, then we set $a = 1$. The opposite extreme case is obtained by setting $a = 0$, that corresponds to total discard of historical information on θ_c. In the basic definition of power priors, the tuning parameter a is either fixed or random, but it does not depend on the available data. In the *dynamic* power prior, on the contrary, a measures the homogeneity between historical and current control data. A natural choice is to consider a measure of agreement between $\pi_c(\cdot|x_c)$ and $\pi_h(\cdot|x_h)$, where $\pi_j(\cdot|x_j)$ are the posterior densities for the control parameter obtained by updating $\pi_j(\cdot)$ with x_j, $j = h, c$. We here consider $\pi_j(\cdot)$ to be $Beta(1, 1)$ densities. Therefore $\theta_c|x_j \sim Beta(\bar{\alpha}_j, \bar{\beta}_j)$, where $\bar{\alpha}_j = 1 + x_j$ and $\bar{\beta}_j = 1 + n_j - x_j$, $j = c, h$. With this purpose, following [23], we first consider a measure based on the Hellinger distance between the two posterior densities, i.e.

$$d[\pi_c(\cdot|x_c), \pi_h(\cdot|x_h)] = \left(1 - \int_{\mathbb{R}} \sqrt{\pi_c(\theta|x_c) \cdot \pi_h(\theta|x_h)}\, d\theta\right)^{\frac{1}{2}}. \tag{3}$$

Then, we define the power prior parameter as the product of two factors: the first is a *static* coefficient $\kappa \in [0, 1]$ that provides an upper limit to the *quantity of information* that we are willing to borrow; the second is a *dynamic* fraction that depends on the *commensurability* between current and historical data, i.e.

$$a(x_c, x_h) = \kappa \cdot (1 - d[\pi_c(\cdot|x_c), \pi_h(\cdot|x_h)]). \tag{4}$$

Table 1 Current data (Rotavirus vaccine example)

Arm	j	n_j	x_j	$\hat{\theta}_j$
Experimental	e	558	415	0.74
Control	c	592	426	0.72

Since $d[\cdot, \cdot]$ is a relative distance and $\kappa \in [0, 1]$, then $a(x_c, x_h) \in [0, 1]$: for a given value of κ, the more compatible information provided by $\pi_c(\cdot|x_c)$ and $\pi_h(\cdot|x_h)$, the larger $a(x_c, x_h)$. Note that, for instance, if we set $\kappa = 1$ the amount of borrowing is fully determined by $(1 - d[\pi_c(\cdot|x_c), \pi_h(\cdot|x_h)])$; conversely, if we set $\kappa < 1$ we impose an upper limit to the fraction to be borrowed. This choice makes sense for instance when $n_h \gg n_c$ and we want to downweight historical prior information so that current data are not overwhelmed. It can be easily checked that under our assumptions (3) becomes

$$d[\pi_c(\cdot|x_c), \pi_h(\cdot|x_h)] = \left(1 - \frac{B\left(\frac{\bar{\alpha}_c+\bar{\alpha}_h}{2}, \frac{\bar{\beta}_c+\bar{\beta}_h}{2}\right)}{\sqrt{B(\bar{\alpha}_c, \bar{\beta}_c) \cdot B(\bar{\alpha}_h, \bar{\beta}_h)}}\right)^{\frac{1}{2}},$$

where $B(\cdot, \cdot)$ is the Beta function.

3 Application

In this section we consider an example described in [19], where a NI study is conducted to compare a pentavalent vaccine (RotaTeq) with a placebo against Rotavirus, both administered together with routine pediatric vaccines. The data are the number of subjects in the two groups who give a positive response to vaccination. Let $\hat{\theta}_j = x_j/n_j$, $j = e, c, h$ denote the response rates.

Table 1 reports current data for both experimental and control arms, whereas Table 2 summarizes data on the control related to four different historical studies that are also combined using a meta-analytic model (*pooled*) as in [19]. In addition, for the sake of the following illustration, we consider the *cumulative* data that are obtained by crude aggregation of the four historical datasets. In the original example, Liu sets $\delta = 0.10$. Here with no loss of generality we consider a stricter NI margin by setting $\delta = 0.03$.

Table 3 reports the values of a computed with Eq. (4) for each single historical study and for cumulative and pooled data. Correspondingly the table shows the bounds L and U of the 0.95-credible intervals and $P(H_1|x_c, x_e)$, i.e. the posterior probability of H_1 computed with respect to (2). Cases $a = 0$ (no borrowing) and $a = 1$ (full borrowing) are also considered for comparison. In Study 1, the degree of borrowing is close to 1 due to the high compatibility between current and histor-

Table 2 Historical data (Rotavirus vaccine example)

Study	n_h	x_h	$\hat{\theta}_h$
1	576	417	0.724
2	111	90	0.811
3	62	49	0.790
4	487	376	0.772
Pooled	483	367	0.759
Cumulative	1236	932	0.754

Table 3 Values of a (with $\kappa = 1$ and $\kappa = 0.8$), bounds of C (with $1 - \gamma = 0.95$) and $P(H_1|x_c, x_e)$ for different historical studies

Study	a	L	U	$U - L$	$P(H_1\|x_c, x_e)$
1	1	−0.023	0.065	0.088	0.989
2	1	**−0.041**	0.058	0.099	0.940
3	1	**−0.033**	0.067	0.100	0.968
4	1	**−0.045**	0.044	0.089	0.901
Pooled	1	**−0.039**	0.050	0.089	0.938
Cumulative	1	**−0.042**	0.041		0.921
	($\kappa = 1$)				
1	0.917	−0.024	0.066	0.090	0.987
2	0.171	−0.029	0.072	0.101	0.978
3	0.331	**−0.030**	0.071	0.101	0.976
4	0.213	**−0.033**	0.064	0.097	0.969
Pooled	0.346	**−0.034**	0.062	0.096	0.964
Cumulative	0.307	**−0.036**	0.056	0.092	0.958
	($\kappa = 0.8$)				
1	0.734	−0.025	0.067	0.092	0.985
2	0.137	−0.029	0.073	0.102	0.977
3	0.265	−0.028	0.074	0.102	0.979
4	0.170	**−0.031**	0.067	0.098	0.973
Pooled	0.277	**−0.031**	0.066	0.097	0.972
Cumulative	0.245	**−0.035**	0.058	0.093	0.962
—	0	−0.027	0.075	0.102	0.981

ical data, both in terms of response rate ($\hat{\theta}_h \approx \hat{\theta}_c$) and study dimension ($n_h \approx n_c$). Conversely, in Study 2 the degree of borrowing is the lowest due to heterogeneity in both sample size and response rate. Study 3 represents an intermediate situation with respect to the previous cases. It is interesting to note that even though the data from Study 4 and the pooled case are very similar in sample size, the values of a are substantially different as a consequence of the difference between the two response rates. Finally, when $n_h > n_c$, as in the cumulative historical data case, the value of a is smaller than in the pooled case as a combined effect of two conflicting determinants: (i) a response rate (slightly) closer to $\hat{\theta}_c$, which is supposed to yield a (slightly) larger a; and (ii) a sample size n_h much larger than n_c, that reduces the commensurability between the two posterior distributions thus yielding a smaller a. This dynamic results in the desired downweighting of historical prior information. As commented in Sect. 2.1 one may be willing to fix an upper limit to the degree of borrowing by setting $\kappa < 1$. For instance, if $\kappa = 0.8$, in Study 1 a reduces to 0.734 with respect to the previous case, which is equivalent to a 26% reduction in the prior sample size (from 576 to $a \cdot n_h = 0.734 \cdot 576 = 423$). See Table 3 for other numerical examples. For a deeper insight in Fig. 1 (top panel) we consider prefixed values of $\hat{\theta}_h$, we compute x_h for each value of n_h from 10 to 1300, we determine the corresponding value of $a(x_c, x_h)$ and plot $a(x_c, x_h)$ as a function of n_h. Circles denote the values of a obtained in Table 3 for the different historical data sets. Note that this plot shows how the concurrent effect of n_h and x_h determines very different steepness of the curves that describe $a(x_c, x_h)$ as a function of n_h and illustrates how in certain cases even very small changes in n_h and x_h produce significant variations in a (as an example compare the values of a corresponding to Studies 2 and 3). As expected the maximum level of compatibility is achieved for $\hat{\theta}_h = \hat{\theta}_c$ and $n_h = n_c$ (solid line) i.e. when $\pi_c(\cdot|x_c) = \pi_h(\cdot|x_h)$. The same conclusion can be drawn by looking at the dotted curve corresponding to $n_h = 592$ in Fig. 1 (bottom panel), where $a(x_c, x_h)$ is now plotted against $\hat{\theta}_h$ for several fixed value of n_h. First of all, note that the maximum value of a is not monotone with n_h. The level of borrowing is due to the combined effect of $\hat{\theta}_h$ and n_h for given values of $\hat{\theta}_c$ and n_c. For fixed values of n_h, the plots of a are symmetric with respect to the values of $\hat{\theta}_h$, i.e. the level of borrowing only depends on the absolute value of the difference $\hat{\theta}_h - \hat{\theta}_c$. Finally, note that in all the empirical studies considered in this example $\hat{\theta}_h > \hat{\theta}_c$.

Let us now comment on the conclusions of the NI test recalling that values of $L < -\delta$ (bold character in Table 3) do not allow to reject the null hypothesis. First of all, if we consider the full borrowing case ($a = 1$) all studies but the first prevent one from rejecting H_0, whereas ignoring historical information ($a = 0$) implies rejection ($L = -0.027 > -0.03 = -\delta$). Secondly, consider the dynamic borrowing case. In Study 1 the high compatibility with the current control data, both in terms of sample sizes and response rates, implies conclusions consistent with the full and the no borrowing cases (regardless of κ). Conversely, results from Study 2 are very different from current control data ($\hat{\theta}_h \gg \hat{\theta}_c$): full borrowing yields a value of L much lower than $-\delta$, whereas, thanks to the small degree of borrowing ($a = 0.171$), one is able to reject the null hypothesis consistently with the total discount case. Similar considerations apply to Study 3, in which $\hat{\theta}_h$ is smaller than in Study 2 and closer to $\hat{\theta}_c$. Due to the

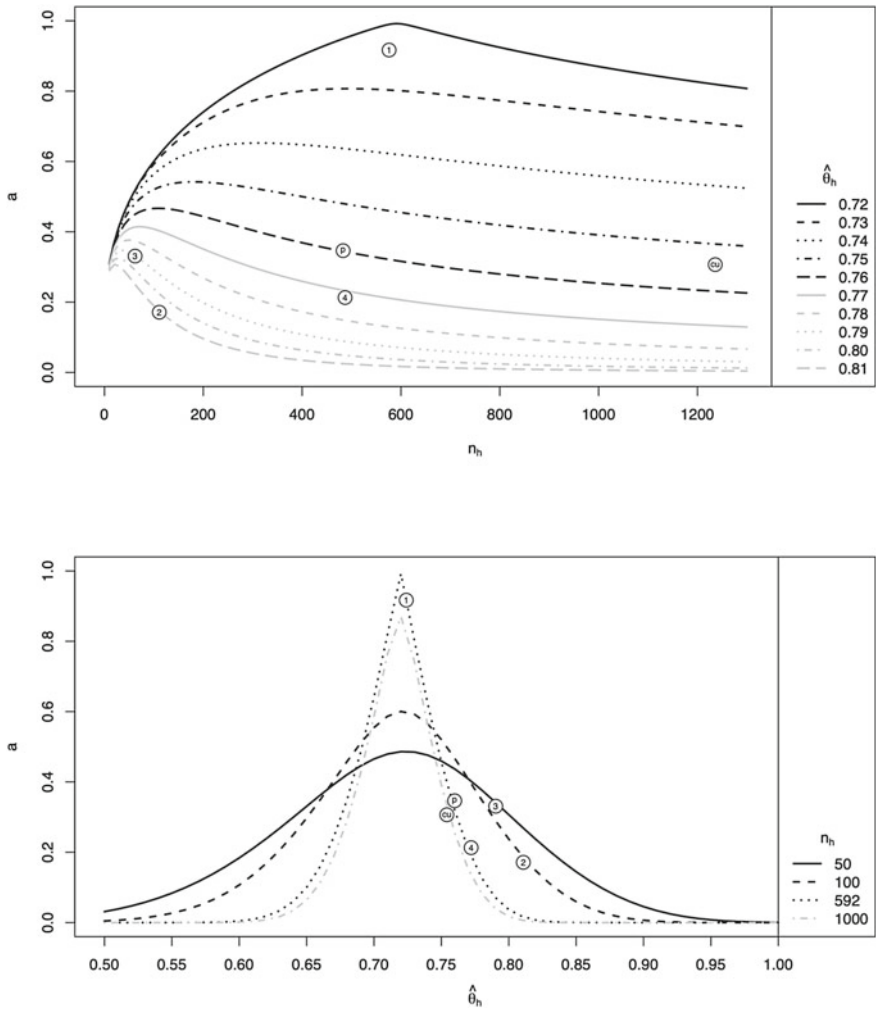


Fig. 1 Behavior of $a(x_c, x_h)$ for increasing values of n_h (top panel) and $\hat{\theta}_h$ (bottom panel). Circles denote the values of a obtained in Table 3 for the different historical data sets (denoted by 1, 2, 3, 4, p = *pooled*, cu = *cumulative*)

larger value of a, H_0 is still accepted for $\kappa = 1$, but rejected for $\kappa = 0.8$. Finally, Studies 4, pooled and cumulative, characterized by $\hat{\theta}_h > \hat{\theta}_c$ and large sample sizes, yield evidence of negative values of θ and determine acceptance of H_0 even for small values of a. For instance, to obtain a value of $L \leq -\delta$ (e.g. $L = -0.029$) in the pooled study case, one should set the quantity of information κ as small as 0.3.

4 Frequentist Type-I Error and Power

Regulatory agencies [2] require new statistical methodologies for the analysis of clinical trials to be evaluated in terms of their frequentist properties, such as type-I error and power. Recall that we here reject the hypothesis $H_0 : \theta \leq -\delta$ vs. the alternative $H_1 : \theta > -\delta$ if $L > -\delta$, where L is the lower bound of the $(1-\gamma)$-credible interval for $\theta = \theta_e - \theta_c$. Note that L is not only a function of the random variables $(\mathbf{X}_e, \mathbf{X}_c)$ but also of the historical data x_h, that we assume to be given. The power function of the test is then defined as $\eta(\theta) = \mathbb{P}[L > -\delta]$, where $\mathbb{P}[\cdot]$ is evaluated with respect to the joint probability mass function of the random variables $(\mathbf{X}_e, \mathbf{X}_c)$ which depends on (θ_e, θ_c). Assessment of type-I error and power requires one to fix specific values of θ, that we denote as θ_d (design value). If we set $\theta_d = -\delta + \xi$, we can express the power function as a function of $\xi \in \mathbb{R}$:

$$\eta(\xi) = \begin{cases} \alpha(\xi) & \xi \leq 0 \\ 1 - \beta(\xi) & \xi > 0 \end{cases} \tag{5}$$

where $\alpha(\xi)$ and $\beta(\xi)$ are the type-I and type-II error functions respectively. The type-I error (size of the test) is $\alpha = \alpha(0)$. We denote as α_s and $1 - \beta_s(\xi)$, $s = 0, a, 1$, the type-I error and the power functions for full, partial, null borrowing respectively.

Let us consider the pooled study of the previous example, in which real data show that $\hat{\theta}_c < \hat{\theta}_h = 0.76$ (setup (a)). For comparison, we also consider two fictitious historical studies (keeping the same study dimension $n_h = 483$) such that $\hat{\theta}_c = \hat{\theta}_h = 0.72$ (setup (b)), $\hat{\theta}_c > \hat{\theta}_h = 0.68$ (setup (c)). For each setup we consider a simulation to obtain the values of $\eta(\xi)$. The plots are shown in Fig. 2 in case of full, dynamic and no borrowing. The steps of the simulation are the following.

1. Specify $x_h, n_h, n_e, n_c, \kappa, \delta, 1-\gamma$.
2. Fix a design value $\theta_c^\star$ for θ_c and generate M values $\tilde{x}_c$ from $Binom(n_c, \theta_c^\star)$.
3. For each $\tilde{x}_c$ compute $a(\tilde{x}_c, x_h)$ according to (4).
4. Draw M values $\tilde{x}_e$ from $Binom(n_e, \theta_e^\star)$, where $\theta_e^\star = \theta_c^\star - \delta + \xi$, with $\xi = 0$ under H_0 and $\xi > 0$ under H_1.
5. Draw B values $\tilde{\theta}_e$ and $\tilde{\theta}_c$ from $\pi_e(\cdot|x_e)$ and $\pi_c^P(\cdot|x_h, x_c)$ and set $\tilde{\theta} = \tilde{\theta}_e - \tilde{\theta}_c$.
6. Compute $\tilde{L}$ as the empirical $(1-\gamma/2)$-quantile of the B values $\tilde{\theta}$.
7. Compute the fraction of $\tilde{L} > -\delta$ and obtain the empirical type-I error (if $\xi = 0$) or the empirical power (if $\xi > 0$).

Figure 2 shows the increasing trend of the power with respect to ξ, for each given combination of setup and borrowing level. Note that, in the partial borrowing case, a is computed according to step 3 of the simulation; the empirical medians over the M simulations are respectively (a) 0.346, (b) 0.770 and (c) 0.412. Table 4 reports the values of α_s, $s = 0, a, 1$. Setup (a), that is based on the pooled real data, is favorable to H_0. Hence a larger a corresponds to smaller α_a, but also to a lower power for each given value of ξ. Conversely, when the contrast between historical and current data goes in the opposite direction, as in setup (c), H_1 is strengthened and, therefore,

Table 4 Type-I error α_s for the three setups and different levels of borrowing, $s = 0, a, 1$. In the partial borrowing case, with $\kappa = 1$, the empirical medians over the M simulations are respectively (a) 0.346, (b) 0.770 and (c) 0.412

α_s	Setup		
	(a)	(b)	(c)
α_1	0.002	0.017	0.081
α_a	0.014	0.023	0.055
α_0	0.025	0.025	0.025

both α_a and $1 - \beta_a(\xi)$ increase with the level of borrowing a. Finally, in case (b) the high compatibility between historical and current control data implies that partial and full borrowing are substantially equivalent in terms of type-I error and power. Not surprisingly, they are both preferable to the no borrowing case, which yields a larger empirical value of α_0 and a lower power.

5 Sample Size Determination

In the previous section we discussed the type-I error and power induced by a prefixed sample size $n = n_e + n_c$ for the current trial (as in the example proposed by [19]) under three alternative historical setups (a), (b), (c). In this section we consider the pre-experimental design problem of selecting the current sample size for both experimental and control arms to be used in the NI test, assuming a fixed value for the NI margin (here $\delta = 0.03$). For simplicity, we consider balanced allocation (i.e. a fixed ratio $r = n_e/n_c = 1$). Given fixed historical data (x_h, n_h) for each setup, we use the power prior in (2) with fraction $a(x_c, x_h)$ obtained from (4) and $\kappa = 1$. Our goal is to find the minimum total sample size n so that the power is at least $1 - \beta^*$, i.e.

$$n^*(\xi) = \min\{n : \ \beta(\xi) \leq \beta^*\} \tag{6}$$

A common choice of β^* is 0.2, so that the power is at least 0.80. The value of α is therefore implied by the fixed value of δ and the selected value of n^*. If the resulting α is not satisfatory (for instance larger than 0.05), one can tune either the value of δ or the threshold β^* (thus selecting a larger n^*). Figure 3 shows the values of α_s and $1 - \beta_s(\xi)$, for $\xi = 0.1$ and $s = 0, a, 1$, as functions of n, computed according to the simulation scheme described in the Sect. 4. Table 5 reports the optimal sample sizes n_s^* under the three setups and the corresponding α_s, $s = 0, a, 1$. Overall, the values of the type-I error are always controlled at a level (smaller than) 0.05, except for setup (c) where α_1 is at most 0.08. In setup (a)—where historical data empower H_0—for each n, $\beta_0(\xi) < \beta_a(\xi) < \beta_1(\xi)$; therefore $n_0^* < n_a^* < n_1^*$ (see Fig. 3, top panel). The opposite relation is observed in setup (c)—which strenghtens H_1—as shown in Fig. 3, bottom panel. Finally, in setup (b)—that is characterized by the strong compatibility

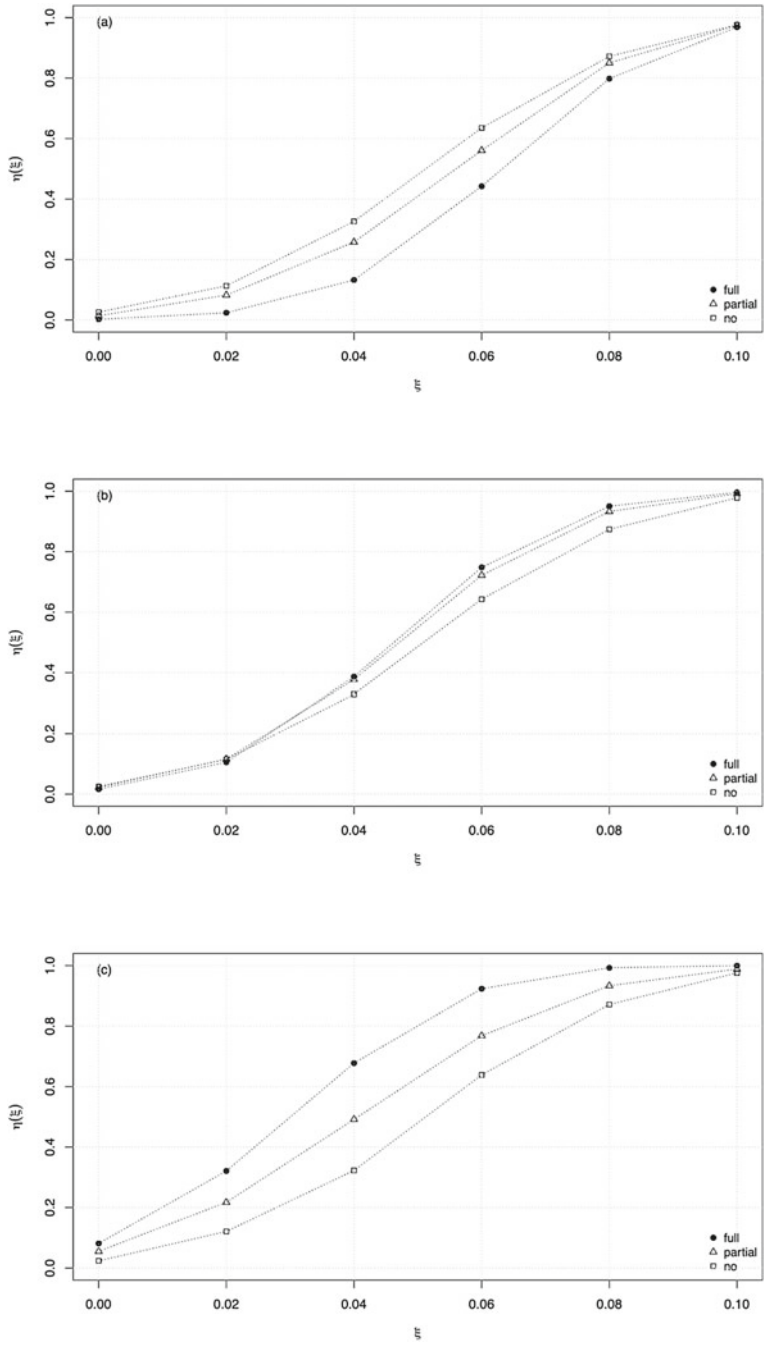


Fig. 2 Power function $\eta(\xi)$ for full, partial and no borrowing under the three scenarios (a), (b), (c). The empirical medians over the M simulations in the partial borrowing case are respectively **a** 0.346, **b** 0.770 and **c** 0.412

Table 5 Optimal sample sizes n_s^* and values of α_s for the three setups (a), (b), (c) and for different levels of borrowing $s = 0, a, 1$

	Setup		
	(a)	(b)	(c)
n_1^*	684	396	256
α_1	0.001	0.015	**0.077**
n_a^*	632	444	326
α_a	0.025	0.026	0.026
n_0^*	582	598	592
α_0	0.025	0.023	0.027

between historical and current control data—n_a^* and n_1^* are relatively close, due to the similar values of $\beta_a(\xi)$ and $\beta_1(\xi)$, whereas $\beta_0(\xi)$ and therefore n_0^* are larger.

6 Conclusions

In this paper we consider a Bayesian test of NI based on a credible interval for $\theta = \theta_e - \theta_c$ and we propose to use a dynamic power prior to build the posterior distribution for the control parameter θ_c, in the spirit of [19]. Specifically, we follow [23] and assume the power prior parameter fraction to be a function of the Hellinger distance between the posterior densities of the control parameter given the current and historical data of the control arms, respectively. The use of the Hellinger distance presents some practical advantages over alternatives approaches. The first is that $d[\cdot, \cdot]$ and, as a consequence, $a = \kappa \cdot (1 - d[\cdot, \cdot])$ is a relative index, i.e. it automatically ranges in $[0, 1]$. Secondly, at least for standard but widely used models, (such as the beta-binomial model considered here) the Hellinger distance is available in closed-form. In more general cases (non-conjugate models), Monte Carlo numerical approximations can be easily implemented. The R [26] code used to perform the the analyses presented in Sects. 3, 4 and 5 is available upon request.

In the remaining of this section we list a few critical issues and some potential areas of further development.

1. *Sample size adjustment*. The Hellinger distance between the two posterior densities in (4) can be highly affected by n_c/n_h. To avoid this, [23] suggests an adjustment to make the likelihoods comparable in terms of sample sizes. In our notation, if $n_h \gg n_c$, which is the most standard case in clinical trials, one should raise the likelihood $f(x_h|\theta_c)$ in $\pi_h(\cdot|x_h)$ to the factor n_c/n_h. Conversely, if $n_c \gg n_h$ the likelihood $f(x_c|\theta_c)$ in $\pi_c(\cdot|x_c)$ should be raised to n_h/n_c. To encompass these two cases, let $\tilde{\pi}_c(\theta_c|x_c) \propto \pi_c(\theta_c) \times [f(x_c|\theta_c)]^{\min\left(1, \frac{n_h}{n_c}\right)}$ and $\tilde{\pi}_h(\theta_c|x_h) \propto \pi_h(\theta_c) \times [f(x_h|\theta_c)]^{\min\left(1, \frac{n_c}{n_h}\right)}$ and let $\tilde{a}$ be the Hellinger distance

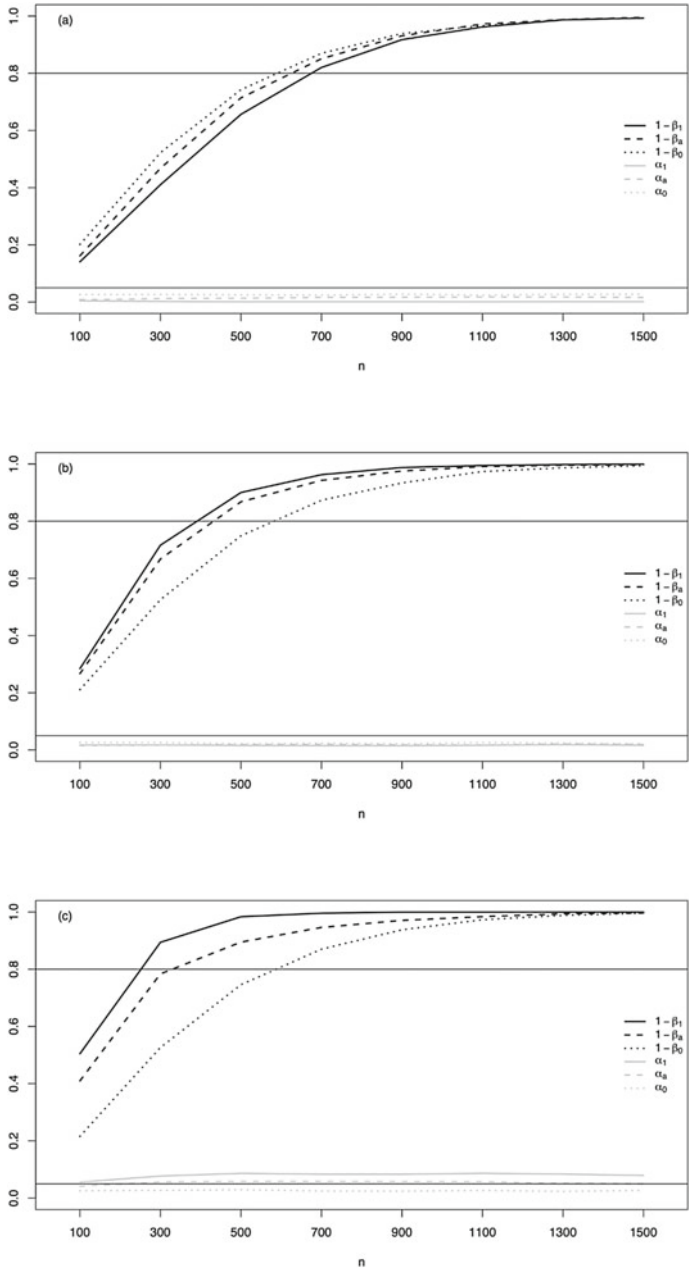


Fig. 3 Values of α_s and $1-\beta_s(\xi)$, for $\xi = 0.1$, as functions of n, for full ($s = 1$), partial ($s = a$) and no ($s = 0$) borrowing under the three historical setups (a), (b), (c)

between $\tilde{\pi}_c(\theta_c|x_c)$ and $\tilde{\pi}_h(\theta_c|x_h)$, i.e.

$$\tilde{a}(x_c, x_h) = \kappa \cdot (1 - d\left[\tilde{\pi}_c(\cdot|x_c), \tilde{\pi}_h(\cdot|x_h)\right]). \tag{7}$$

It can be checked that after this sample size adjustment the distance $d[\cdot, \cdot]$ essentially depends on the difference between $\hat{\theta}_h$ and $\hat{\theta}_c$ only. Hence, the sample size adjustement implies overlooking of one relevant component of the difference between the two distributions. Furthermore, the unadjusted Hellinger distance is able to account for the global difference between distributions. As a consequence, for instance when $n_h \gg n_c$ the risk of using overwhelming historical data is automatically handled (compare dotted and dashed-dotted lines in Fig. 1); otherwise an upper limit can be externally fixed by setting κ as suggested in Sect. 2.1. In addition, downweighting π_c in case $n_c \gg n_h$ would be justified only if we treated current and historical information with the same importance, which is not the case.

2. *NI margin*. The choice of the non-inferiority margin is of crucial importance (see [3, 7] for guidelines). In this paper we consider the NI margin prefixed according to the specific problem of interest [4, 18, 19, 22]. For alternative approaches see [10]. Note that the values of δ and n determine the α level of the NI test. Therefore this choice must be carefully evaluated when control of type-I error is of concern.
3. *Predictive vs conditional approach*. Preposterior analyses of Sect. 4 (frequentist type-I error and power for fixed n) and of Sect. 5 (sample size determination) are here perfomed using the sampling distribution of the data $f(\mathbf{x}_e, \mathbf{x}_c|\theta_e, \theta_c)$ for a fixed design value θ_d. As an alternative to the frequentist power function, several authors advise the use the Bayesian power [4, 18, 25]. This is still defined as the probability of the event $L(\mathbf{X}_e, \mathbf{X}_c) > -\delta$ but it is now computed with the prior predictive distribution $m(\mathbf{x}_e, \mathbf{x}_c)$ of the data, defined as $m(\mathbf{x}_e, \mathbf{x}_c) = \int f(\mathbf{x}_e, \mathbf{x}_c|\theta_e, \theta_c) \cdot \pi_d(\theta_e, \theta_c) d\theta_e d\theta_c$, where $\pi_d(\theta_e, \theta_c)$ is called design prior. The replacement of the sampling distribution $f(\mathbf{x}_e, \mathbf{x}_c|\theta_e, \theta_c)$ with $m(\mathbf{x}_e, \mathbf{x}_c)$ makes it possible to avoid the dependence of the test power on a specific θ_d and to take into account uncertainty on the design value. For the sake of simplicity we here focus only on the standard frequentist power, even though extensions to the Bayesian power are straightforward.
4. *Random historical data*. In our preposterior analyses we assume historical data to be fixed. Alternatively, one might also take into account potential randomness of X_h and study its impact on the performances in terms of type-I error and power.
5. *Allocation of experimental units*. In this paper we consider a fixed allocation ratio (e.g. $r = n_e/n_c = 1$). This reduces the design problem to the selection of n. More refined allocation rules will be hopefully explored in future research.

References

1. Brutti, P., De Santis, F., Gubbiotti, S.: Bayesian-frequentist sample size determination: a game of two priors. Metron **72**(2), 133–151 (2014)
2. CDHR/FDA: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Guidance for industry and FDA staff (2010)
3. CDER-CDBR/FDA: Non-Inferiority clinical trials to establish effectiveness. Guidance for industry (2016)
4. Chen, M.H., Ibrahim, J.G., Lam, P., Yu, A., Zhang, Y.: Bayesian design of noninferiority trials for medical devices using historical data. Biometrics **67**, 1163–1170 (2011)
5. De Santis, F.: Power priors and their use in clinical trials. Am. Stat. **60**(2), 122–129 (2006)
6. De Santis, F.: Using historical data for Bayesian sample size determination. J. R. Stat. Soc. Ser. A. **170**, 95–113 (2007)
7. EMA Guideline on the choice of non-inferiority margin. EMEA/CPMP/EWP/2158/99 (2005)
8. Gamalo, M.A., Wu, R., Tiwari, R.C.: Bayesian approach to noninferiority trials for proportions. J. Biopharma. Stat. **21**(5), 902–919 (2011)
9. Gamalo, M.A., Wu, R., Tiwari, R.C.: Bayesian approach to non-inferiority trials for normal means. Stat. Methods Med. Res. **25**(1), 221–240 (2016)
10. Gamalo-Siebers, M., Gao, A., Lakshminarayanan, M., Liu, G., Natanegara, F., Railkar, R., Schmidli, H., Song, G.: Bayesian methods for the design and analysis of noninferiority trials. J. Biopharm. Stat. **26**(5), 823–841 (2016)
11. Gamalo, M.A., Tiwari, R.C., LaVange, L.M.: Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. Pharmaceut. Stat. **13**, 25–40 (2014)
12. Gravestock, I., Held, L.: Adaptive power priors with empirical Bayes for clinical trials. Pharma. Stat. **16**, 349–360 (2017)
13. Gravestock, I., Held, L.: Power priors based on multiple historical studies for binary outcomes. Biometrical J. **61**, 1201–1218 (2019)
14. Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J.: Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics **67**, 1047–1056 (2011)
15. Hobbs, B.P., Sargent, D.J., Carlin, B.P.: Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. **7**(3), 639–674 (2014)
16. Harun, N., Liu, C., Kim, M.O.: Critical appraisal of Bayesian dynamic borrowing from an imperfectly commensurate historical control. Pharma. Stat. **19**, 613–625 (2020)
17. Ibrahim, J.G., Chen, M.H.: Power prior distributions for regression models. Stat. Sci. **15**, 46–60 (2000)
18. Li, W., Chen, M.H., Wang, X., Dey, D.K.: Bayesian design of non-inferiority clinical trials via the Bayes factor. Stat. Biosci. **10**, 439–459 (2018)
19. Liu, G.F.: A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. Pharm Stat. **17**, 61–73 (2018)
20. Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D.J.: Summarizing historical information on controls in clinical trials. Clin. Trials. **7**, 5–18 (2010)
21. Nikolakopoulos, S., van der Tweel, I., Roes, K.C.B.: Dynamic borrowing through empirical power priors that control type I error. Biometrics **74**, 874–880 (2018)
22. Osman, M., Ghosh, S.K.: Semiparametric Bayesian testing procedure for noninferiority trials with binary endpoints. J. Biopharm. Stat. **21**(5), 920–937 (2011)
23. Ollier, A., Morita, S., Ursino, M., Zohar, S.: An adaptive power prior for sequential clinical trials—application to bridging studies. Stat. Methods Med. Res. **29**(8), 2282–2294 (2020)
24. Pan, H., Yuan, Y., Xia, J.: A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. Appl. Stat. **66**, 979–996 (2017)
25. Psioda, M.A., Ibrahim, J.G.: Bayesian clinical trial design using historical data that inform the treatment effect. Biostatistics **20**(3), 400–415 (2019)

26. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2020)
27. Ventz, S., Trippa, L.: Bayesian designs and the control of frequentist characteristics: a practical solution. Biometrics **71**, 218–226 (2015)
28. Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J.G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., Thompson, L.: Use of historical control data for assessing treatment effects in clinical trials. Pharm. Stat. **13**(1), 41–54 (2014)
29. Wang, F., Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Stat. Sci. **17**(2), 193–208 (2002)

A Graphical Approach for the Selection of the Number of Clusters in the Spectral Clustering Algorithm

Cinzia Di Nuzzo and Salvatore Ingrassia

Abstract The selection of the number of clusters K in a data set is a fundamental issue in data clustering. In this paper we consider the problem of the selection of the optimal number of clusters in the spectral clustering algorithm. To this end, we propose a multi-graphic method that takes into account geometric characteristics derived from the similarity matrix and from the Laplacian embedding among data. Our approach is supported by some mathematical properties developed in the context of spectral clustering theory. Finally, the performance of our proposal is illustrated on ground of some numerical studies based on both synthetic and real datasets.

Keywords Spectral clustering · Laplacian embedding · Number of clusters

1 Introduction

Spectral clustering methods have become very popular for finding non-convex clusters of data, see e.g. [10, 16]. These methods are based on the graph theory, where the data are represented by the vertices in an undirected graph and the edges are weighted by the similarities. In particular, the spectral approach is based on the properties of the pairwise similarity matrix coming from a suitable kernel function and the clustering problem is reformulated as a graph partition problem.

The spectral clustering algorithm does not work directly on the raw data but works on the embedded data in a suitable feature space having a smaller (and even a very smaller) dimension than the space of the original data. This implies that this algorithm is a handy approach for handling high-dimensional data.

Determining the number of clusters is a fundamental issue in data clustering, see [6, 11, 13]. Here, we consider a significant problem concerning the selection of the

C. Di Nuzzo (✉)
Department of Statistics, Sapienza University of Rome, Rome, Italy
e-mail: cinzia.dinuzzo@uniroma1.it

S. Ingrassia
Department of Economics and Business, University of Catania, Catania, Italy
e-mail: s.ingrassia@unict.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_3

number of clusters in the spectral clustering algorithm. In this framework, several authors have proposed criteria for estimating the number of groups based on the analysis of the eigenvectors of the Laplacian matrix, see [7, 17]; however, these criteria do not always provide good results, so estimating the number of groups in spectral clustering is still an open problem. Like in other unsupervised clustering procedures, the spectral clustering algorithm requires the user to specify the number of clusters K as an input value. In the spectral clustering literature, some approaches have been proposed. The most famous criterion is based on the analysis of the maximum eigengap between two consecutive eigenvalues, which is well explained by perturbation theory and graph theory; anyway, in the analysis of real datasets this approach does not provide good results, see [16].

Another approach for the selection of the number of clusters is based on the minimization of a suitable cost function built on rotated eigenvectors of the Laplacian matrix, see [17]; more recently, [7] proposed a criterion based on the multimodality of the Dip test statistics computed on the eigenvectors of the Laplacian matrix. Unfortunately, these criteria work quite well on simulated data but in general, do not provide good results when we have to face with real datasets.

In this paper, we follow a different approach for the selection of the number of clusters K taking into account different criteria that derive from well-established mathematical properties in the spectral clustering theory. In particular, we suggest a criterion based on a joint graphical analysis of different geometrical ingredients: the similarity matrix, the eigengap values, and the shape of the eigenvectors after the Laplacian embedding. Our approach is illustrated on the ground of some numerical studies based on both simulated and real datasets.

The rest of the paper is organized as follows: in Sect. 2 we summarize the spectral clustering method; in Sect. 3 we introduce our proposal for the selection of the number of clusters; Sect. 4 presents numerical studies based on synthetic and real datasets. Finally, in Sect. 5, we provide some concluding remarks and give ideas for future research.

2 Spectral Clustering

Let $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of points in $\mathscr{X} \subseteq \mathbb{R}^p$. In order to group the data V in K cluster, the first step concerns the definition of a symmetric and continuous function $\kappa : \mathscr{X} \times \mathscr{X} \to [0, \infty)$ called the *kernel function*. Afterwards, a *similarity matrix* $W = (w_{ij})$ can be assigned by setting $w_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$, for $\mathbf{x}_i, \mathbf{x}_j \in \mathscr{X}$; in particular, in spectral clustering algorithms, a quite popular choice is the *Gaussian kernel* given by

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\epsilon) \tag{1}$$

for some fixed parameter $\epsilon > 0$.

The choice of the kernel function in spectral clustering algorithms is crucial because it affects the entire data structure in the graph, and consequently, the struc-

ture of the Laplacian and its eigenvectors. An optimal kernel function should lead to a similarity matrix W having (as much as possible) diagonal blocks: in this case, we get well-separated groups and we are also able to understand the number of groups in that data set by counting the number of blocks. In the Gaussian kernel (1) the main problem concerns the choice of the scale parameter ϵ. To this end, the following kernel function is proposed in [17] based on a local scaling parameter ϵ_i for each $\mathbf{x}_i$

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\epsilon_i \epsilon_j}\right) \tag{2}$$

with $\epsilon_i = \|\mathbf{x}_i - \mathbf{x}_h\|$, where $\mathbf{x}_h$ is the h-th neighbor of point $\mathbf{x}_i$ (similarly for ϵ_j); this approach has been called *self-tuning* in [17]. The kernel (2) leads to a similarity matrix that depends on the pairwise proximity between the points. However, despite the name *self-tuning*, the approach is not completely automatic because we have to select the number h of neighbors of the point $\mathbf{x}_i$. To this end, the value $h = 7$ is suggested in [17]. Unfortunately, numerical studies show that this choice cannot be adopted in general.

We remark that other kernel functions have been proposed in the literature, see e.g. [7] and [18]; anyway, to fix the ideas and for simplicity, in the following, we consider the self-tuning kernel function (2).

Once the similarity matrix W (based on the kernel function) has been computed, we introduce the *normalized graph Laplacian* as the matrix $L_{\text{sym}} \in \mathbb{R}^{n \times n}$ given by

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}, \tag{3}$$

where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is the *degree matrix*, d_i is the *degree* of the vertex $\mathbf{x}_i$ defined as $d_i = \sum_{j \neq i} w_{ij}$ and I denotes the $n \times n$ identity matrix. The Laplacian matrix L_{sym} is positive semi-definite with n non-negative eigenvalues. For a fixed $K \ll n$, let $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K\}$ be the eigenvectors corresponding to the smallest K eigenvalues of L_{sym}. Then the *normalized Laplacian embedding in the K principal subspace* is defined as the map $\boldsymbol{\Phi}_{\boldsymbol{\Gamma}} : \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \to \mathbb{R}^K$ given by

$$\boldsymbol{\Phi}_{\boldsymbol{\Gamma}}(\mathbf{x}_i) = (\gamma_{1i}, \ldots, \gamma_{Ki}), \quad i = 1, \ldots, n,$$

where $\gamma_{1i}, \ldots, \gamma_{Ki}$ are the i-th components of $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$, respectively. In other words, the function $\boldsymbol{\Phi}_{\boldsymbol{\Gamma}}(\cdot)$ maps the data from the input space $\mathscr{X}$ to a *feature space* defined by the K principal subspace of L_{sym}. Afterwards, let $\boldsymbol{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')$ be the $n \times K$ matrix , the embedded data in the feature space, where $\mathbf{y}_i = \boldsymbol{\Phi}_{\boldsymbol{\Gamma}}(\mathbf{x}_i)$ for $i = 1, \ldots, n$. The embedded data $\boldsymbol{Y}$ are clustered according to some clustering procedure, usually, the k-means algorithm is taken into account. The spectral clustering algorithm is summarized in Algorithm 1, see also [12]. In this algorithm, we can distinguish two main parts: Steps 1–5 concern the construction of the feature space, and Step 6 concerns the clustering process.

Algorithm 1 Spectral Clustering Algorithm

Input: dataset V, number of clusters K, kernel function κ (and related parameter(s)).

1. Compute the similarity matrix W.
2. Compute the normalized graph Laplacian L_{sym} based on W.
3. *Eigendecomposition*: compute the K smallest eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_K$ and consider the eigenvectors $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K$ corresponding to the eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$ of L_{sym}.
4. *Embed the data in the K-th principal subspace*: compute $\Phi_{\boldsymbol{\Gamma}}(\mathbf{x}_i)$ according to the *normalized Laplacian embedding* and build the matrix $\boldsymbol{Y}$.
5. *Normalization:* Normalize the rows of $\boldsymbol{Y}$ to have unit length yielding $\boldsymbol{X} \in \mathbb{R}^{n \times K}$, such that

$$X_{ij} = \frac{Y_{ij}}{\left(\sum_j Y_{ij}^2\right)^{1/2}}.$$

6. *Data Clustering*: Run a clustering algorithm on the matrix $\boldsymbol{X}$.

Output: the data clustering $C_1, \ldots, C_K$.

We remark that the raw data V can be numerical, text, categorical, or have a more complex structure. Moreover, Step 4 concerns dimensionality reduction of the data in a K-dimensional space through the Laplacian embedding and this is particularly significant when the dimension of the raw data is very large. Furthermore, Step 5 uses an additional row normalization step because this normalization has good noise reduction effects and the eigenvectors computations are more stable numerically, guaranteeing that the eigenvectors will be orthogonal, see [10].

As regards the clustering on the embedded data, we note that usually in Step 6 of Algorithm 1, the mapped data are clustered according to the k-means algorithm. Alternatively, clustering procedure based on Gaussian mixtures have been proposed which allow for more flexible shapes, see [1] and [2].

Finally, it is worth noting that fuzzy approaches to spectral clustering have also been analyzed in the literature, see [3, 5, 14].

On the selection of the number of clusters. Given a dataset V, according to Algorithm 1, the input quantities to be selected in the spectral clustering algorithm are the number of clusters K and the functional form of the kernel function κ (and its parameter(s) as well, for example, the radius ϵ in (1) or the proximity parameter h in (2)). In the following, these parameters will be referred to as the *hyperparameter* of the model. To this end, some approaches and related rules of thumb for the choice of the parameters have been proposed in the literature, but there is no unique criterion that can be adopted in general. In the following, we summarize the main ideas proposed in the literature.

In the spectral clustering approach, if the data are well-separated then the underlying geometric properties of the spectral algorithm allow us to easily select the number of groups. However, in real datasets, the data are not always well separated, and therefore, we can try to resort to other strategies for choosing K.

In this framework, the analysis of the eigenvalues of the Laplacian matrix L_{sym} (3) provides a heuristic method to select the number of clusters K. In the ideal case, the multiplicity of the eigenvalue 0 of L_{sym} is equal to the number of connected components of the graph, since the matrix L_{sym} is a block matrix and the number of blocks is equal to the number of connected components, that is, it is equal to the number of clusters K. The heuristic method selects the number K of groups such that $K = \max_k |\lambda_k - \lambda_{k+1}|$; we remark that this method is also explained by the perturbation theory, see [16]. Nevertheless, if the groups are not well separated, the first eigenvalues are not exactly 0, but they deviate from 0: for this reason, the eigengap method, in general, is not an effective approach in practice.

Another way to select the number K is to consider some theoretical results given in [4, 15]. In fact, in the case of finite samples, [15] gives sufficient conditions under which the spectral clustering maps well-separated data points to approximately orthogonal regions in the embedding space. In particular, [15] proved that under suitable conditions, the embedding has an orthogonal cone structure (OCS). One important consequence is that, when there is little overlap between mixture components with respect to the kernel, the data will concentrate in tight spikes about the axes and therefore, the Laplacian map is relevant to clustering. Thus, [15] and [4] allow us to deduce that, if the embedded data assume a cones structure, then the number of clusters K is the number of the cones/tight spikes in the feature space; in particular, the narrower and more separated spikes, the better results will be provided by the spectral clustering algorithm.

A further criterion is related to the alignment of the eigenvectors, see [17]. As matter of fact, to obtain good clustering results, the eigenvectors must be as much aligned as possible to the indicator vectors and [17] proposed to find the optimal number of groups by trying to align the eigenvectors to the indicator vectors minimizing a cost function, as much as possible.

Finally, the most recent approach has been presented in [7], where the idea is to select K involving an eigenvector distribution analysis, in particular, this method examines the multimodality of the eigenvectors by means of a Dip statistical test.

However, many numerical studies show that, in general, none of these proposals is an effective and general criterion for the selection of the number of clusters. Moreover, as far as the selection of the proximity parameters is concerned, only rules of thumb are proposed in the literature that cannot be assumed as general rules.

3 A Multi-graphical Approach to Select the Number of Clusters

Our approach for the selection pf the proximity parameter h in the kernel function (2) and the number K of clusters, comes from the three following remarks about existing literature:

1. there is a relation between the choice of the proximity parameter(s) in the kernel function and the selection of the number K of clusters and therefore we cannot analyze these two quantities separately;
2. there is no criterion for the choice of the number K of clusters which dominates the other ones and therefore the number of groups can be selected by taking into account some criteria at the same time.

To begin with, we remark that despite this difficulty, the information provided by the geometric features of the Laplacian matrix L_{sym} is useful for this aim. Since the choice of the kernel function affects the entire data structure, our idea is to provide the parameters selected from a joint analysis of three main characteristics: the number of blocks of the similarity matrix W, the maxima values of the eigengaps between two consecutive eigenvalues, and the number of spikes in the embedded data space.

Assume that the dataset V consists of n units. Since we cannot explore all possible values of $h \in \{1, 2, \ldots, n\}$, first we select a subset $\mathscr{H} \subseteq \{1, 2, \ldots, n\}$. For all $h \in \mathscr{H}$ in the self-tuning kernel (2), we take into account: the plot of the similarity matrix W as greyscale, the plot of the eigengap values and the scatter plot of the mapped data in the feature space; afterward, we select the parameters according to the following steps:

1. first, we check if the similarity matrix W highlights a clear diagonal block structure and/or the scatter plot of the embedded data in the feature space shows an aligned or star structure as much as possible;
2. if the similarity matrix W and the embedded data do not show a clear diagonal block structure or orthogonal cone structure, we look at the eigengap plot: if this plot shows a unique maximum eigengap, then K is selected according to this maximum;
3. otherwise, if the multiple maxima emerge from the eigengap analysis in Step 2, we analyze the plots of the embedded data for different K according to the local maxima of the eigengap and select the number of clusters to be not smaller than the number of spikes in the plot of the embedded data.

To summarize, in our proposal we first consider different values of the proximity parameter h in the self-tuning kernel function (2), and subsequently, we select the values of h and K so that the clearest clustering structure emerges from the joint analysis. For example, as h varies, the number of clusters clearly emerges from the blocks of the similarity matrices and/or from the number of spikes in the feature space.

In addition, it is worth noting in the spectral clustering method, the number of clusters is related to the number of the eigenvectors of the Laplacian matrix, so that choosing the number of clusters is equivalent to selecting the dimension of the reduced embedded data (i.e. the number of columns of the Y matrix) to be clustered according to the k-means algorithm. For this reason, we cannot consider common cluster validity indices.

Table 1 Summary information about the datasets

Dataset	Type	Units	Variables	Classes	Class distribution (%)
Skew-t	Synt	360	2	4	16.66, 22.22, 33.33, 27.77
Circle&Squares	Synt	237	2	3	42.19, 34.60, 23.12
Anuran calls	Real	1052	22	3	44.87, 29.47, 25.66

4 Numerical Studies

The approach described in the previous section is now illustrated by numerical examples considering both synthetic and real datasets. Statistical summaries are presented in Table 1. As for the set $\mathscr{H}$ of the proximity parameters introduced before, here we considered the following percentages 1, 5, 10, and 20% of the number of observations n of the dataset.

4.1 Simulation Study: Skew-t Data

A short simulation study is based on random observations from the skew-*t* distribution, see [8, 9]. For simplicity we have selected only two variables; however, we point out that the results presented here are similar to cases with more variables. The *data* consists of $n = 360$ units, $p = 2$ variables and $K = 4$ classes. The groups have been generated according to the following parameters: (a) $n_1 = 60$ random observations, mean vector: $\mu_1 = (0.5, 0.5)$, scale matrix: $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, skewness vector: $\alpha_1 = (1, 3)$, and $\nu_1 = 5$ degrees of freedom; (b) $n_2 = 80$, $\mu_2 = (2, 3)$, $\Sigma_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $\alpha_2 = (5, 4)$, and $\nu_2 = 10$; (c) $n_3 = 120$, $\mu_3 = (5, -2)$, $\Sigma_3 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $\alpha_3 = (8, 6)$, and $\nu_3 = 10$; (d) $n_4 = 100$, $\mu_4 = (1, -6)$, $\Sigma_3 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$, $\alpha_4 = (5, 9)$, and $\nu_4 = 4$. The dataset is shown in Fig. 1.

Numerical results are illustrated in Fig. 2. In Fig. 2a the plot of the similarity matrix W in greyscale is shown to visualize the number of blocks; in Fig. 2b, the first eight eigengaps between two consecutive eigenvalues are shown; in Fig. 2c the scatter plot of the embedded data in the feature space is presented (here only the first three eigenvectors have been plotted). The number of groups is clearly highlighted by the number of blocks in the similarity matrix (Fig. 2a). As for the eigengap values,

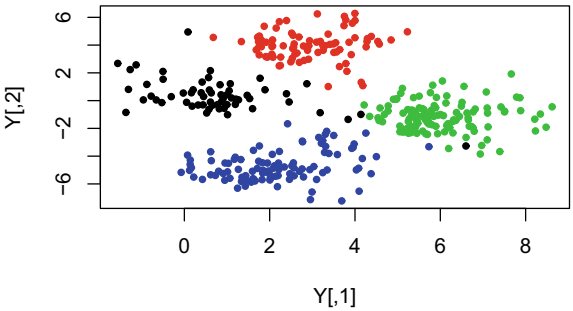


Fig. 1 Scatter plot of the Skew-t data

Table 2 Skew-t data. Accuracy and ARI for different parameters of the self-tuning kernel function

Parameters value	KM	
	Accuracy	ARI
$h = 4$	0.975	0.9426
$h = 18$	0.9722	0.928
$h = 36$	0.975	0.9365
$h = 72$	0.9722	0.9330

Figs. 2b show the right number of clusters for $h = 4$, 18, and 36; while for $h = 72$ the largest eigengaps provides quite similar values for $K = 3$ and $K = 4$. Figure 2c show that the number of spikes in the embedded space is 4. As regards the choice of the proximity parameter h, the first 3 plots of Fig. 2c ($h = 4$, 18, 36) show more separated groups with respect to the case shown in the last row. For sake of completeness, in Table 2 we list the accuracy and ARI values for the Spectral clustering algorithm using k-means algorithm in Step 6 of Algorithm 1.

4.2 Synthetic Data: Circles&Squares Dataset

The *Circles&Squares dataset* consists of $n = 237$ units, $p = 2$ variables and $K = 3$ classes. The dataset is included in the `speccalt R` package and it is shown in Fig. 3.

Numerical results are illustrated in Fig. 4. Figure 4a shows the plot of the similarity matrix W in greyscale to visualize the number of blocks; Fig. 4b shows the first eight eigengaps between two consecutive eigenvalues and finally Fig. 4c presents the scatter plot of the embedded data in the feature space (here only the first three eigenvectors have been plotted).

The number of groups is highlighted by the number of blocks in the similarity matrix (see Fig. 4a) and this is very evident in the cases $h = 12$, $h = 24$, and $h = 47$. As for the eigengap values, Fig. 4b show that the right number of clusters $K = 3$

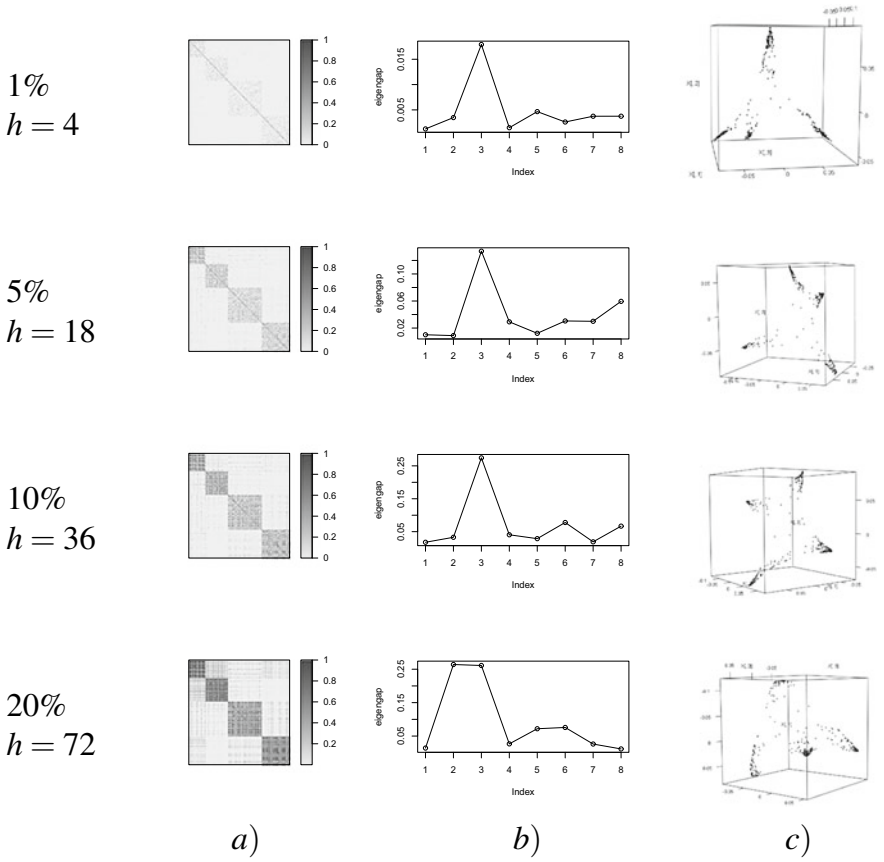


Fig. 2 Spectral clustering features of the Skew-t data varying the h parameter in the self-tuning kernel (2). **a** Similarity matrix W. **b** Eigengap values. **c** Embedded data

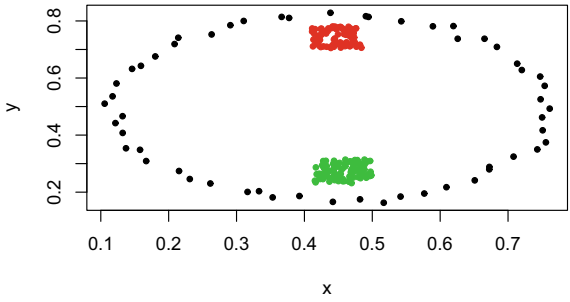


Fig. 3 Scatter plot of the Circle&Squares dataset

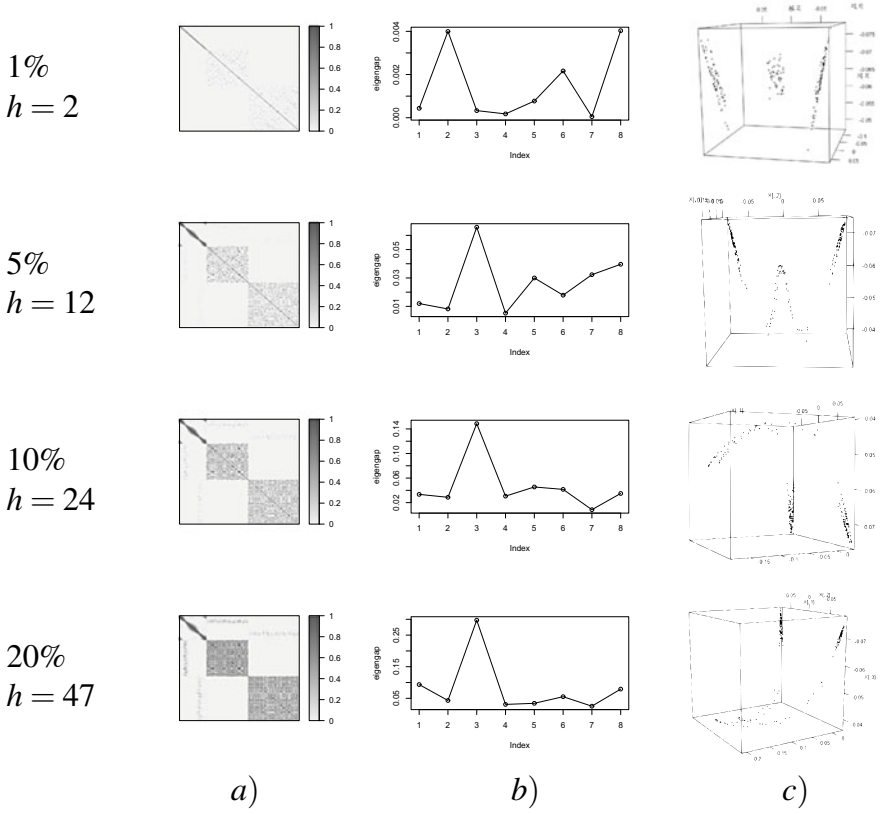


Fig. 4 Spectral clustering features of the Circle&Squares dataset varying the h parameter in the self-tuning kernel (2). **a** Similarity matrix W. **b** Eigengap values. **c** Embedded data

is highlighted only for $h = 2$. Finally, Fig. 4c shows that the number of spikes in the embedded space is 3 and it is especially clear when $h = 2$ and $h = 12$. As for the choice of the proximity parameter h is concerned, we remarked above that the maximum eigengap agrees with the number of blocks and with the number of directions is $h = 2$. As a matter of fact, in this case, we get very separated and fairly aligned directions in the embedded data.

In Table 3 we report the accuracy and ARI values for the spectral clustering algorithm using k-means in the last step of Algorithm 1 varying h. Moreover, the analysis of the misclassified points shows that they are close to the points belonging to the squares in Fig. 3.

Table 3 Circle&Squares dataset. Accuracy and ARI for different parameters of the self-tuning kernel function

Parameters value	KM	
	Accuracy	ARI
$h = 2$	0.9958	0.9893
$h = 12$	0.9831	0.9582
$h = 24$	0.9620	0.9045
$h = 47$	0.9536	0.8828

Table 4 Anuran Calls dataset. Accuracy and ARI for different parameters of the self-tuning kernel function

Parameters value	KM	
	Accuracy	ARI
$h = 11$	0.9639	0.8986
$h = 53$	0.9629	0.8977
$h = 105$	0.9619	0.9861
$h = 210$	0.9600	0.8889
$h = 526$	0.9572	0.8847

4.3 Real Data: Anuran Calls Dataset

The dataset concerns acoustic features extracted from syllables of anuran (frogs) calls. It is a multilabel dataset with three columns of labels: 4 groups for the family, 8 groups for the genus, and 10 groups for the species labels, see https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29. The original dataset consists of 7195 units and 22 attributes; for the scope of the present paper, here we considered a reduced version with 1052 units and three classes that represent three different families.

The numerical results are summarized in Fig. 5. From Fig. 5a we can see that three blocks are well highlighted for $h = 526$. Moreover, the plot of the mapped data Fig. 5c shows clearly three spikes. The same result is confirmed by the analysis of the eigengaps, see Fig. 5b. Therefore, we can state that the number of clusters is $K = 3$. As for the parameter selection in the self-tuning kernel function, we note that the first row has more separated spikes in the feature space, thus we set $h = 11$; but we point out that also the other values provide very separated cones in the feature space. In order to confirm the correctness of our approach, we reported the Accuracy and ARI values in Table 4 for different h.

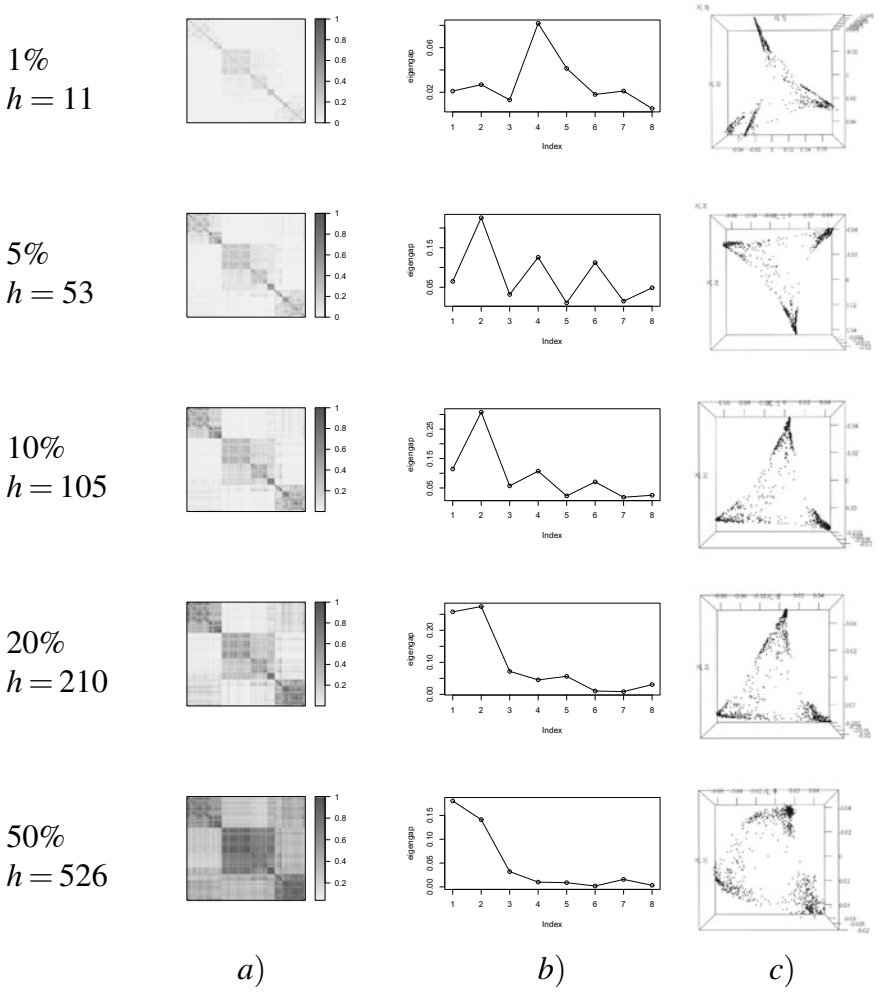


Fig. 5 Spectral clustering features of the Anuran Calls dataset varying the h parameter in the self-tuning kernel (2). **a** Similarity matrix W. **b** Eigengap values. **c** Embedded data

5 Conclusions

In this paper, we have presented a new approach for the selection of the number K of groups in the spectral clustering algorithm and the performance of our proposal has been illustrated by means of numerical studies based on both synthetic and real datasets.

In particular, we propose to select some candidates for K in such a way that a good selection emerges from a joint analysis of three main features: the eigengap values, the number of the blocks of the similarity matrix and the number of the spikes given out by the first eigenvectors. Moreover, we remarked that the choice of the proximity parameter(s) in the kernel function affects the selection of the number K of clusters and therefore we cannot analyze these two quantities separately. Thus, the best way to select it is to evaluate the number of blocks in the similarity matrix and the number of spikes in the feature space, trying to reach an agreement between the number of blocks, the number of eigenvector directions and the eigengap candidates.

Finally, despite the many advantages of our approach, we also want to point out that the computation of the eigengap needs to set a maximum value of clusters. In our experiments, we have chosen $K_{\max} = 10$. Moreover, due to the 3D representation, the spikes in the embedded space are not always well visualized, because sometimes some clusters can be hidden by another cluster. This provides ideas for future research.

References

1. Di Nuzzo, C., Ingrassia, S.: A mixture model approach to spectral clustering and application to textual data. Stat. Methods Appl. https://doi.org/10.1007/s10260-022-00635-4 (2022)
2. Di Nuzzo, C.: Model selection and mixture approaches in the spectral clustering algorithm. Ph.D. Thesis, Economics, Management and Statistics, University of Messina (2021)
3. Feng Z., Hanqiang, L., Licheng, J.: Spectral clustering with fuzzy similarity measure. Digit. Signal Process. **21**(6), 701–709. ISSN 1051-2004 (2011)
4. Garcia Trillos, N., Hoffman, F., Hosseini, B.: Geometric structure of graph Laplacian embeddings. arXiv preprint arXiv:1901.10651 (2019)
5. Hanqiang, L., Feng, Z., Licheng, J.: Fuzzy spectral clustering with robust spatial information for image segmentation. Appl. Soft Comput. **12**(11), 3636–3647, ISSN 1568-4946 (2012)
6. Hennig, C.: Cluster validation by measurement of clustering characteristics relevant to the user. In: Skiadas, C.H., Bozeman, J.R. (eds.) Data Analysis and Applications 1 (2019)
7. John, C.R., Watson, D., Barnes, M.R., Pitzalis, C., Lewis, M.J.: Spectrum: fast density-aware spectral clustering for single and multi-omic data. Bioinformatics **36**(4), 1159–1166 (2019)
8. Lee, S.X., McLachlan, G.J.: Model-based clustering and classification with non-normal mixture distributions. Stat. Methods Appl. **22**(4), 427–454 (2013)
9. Lee, S.X., McLachlan, G.J.: On mixtures of skew normal and skew t-distributions. Adv. Data Anal. Classif. **7**(3), 241–266 (2013)
10. Meila, M.: Spectral clustering. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds) Handbook of Cluster Analysis. Chapman and Hall/CRC (2015)
11. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**, 159–179 (1985)
12. Ng, A.Y., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. Adv. Neural Inf. Process. Syst. **14** (2002)
13. Pappadá, R., Pauli, F., Torelli, N.: Assessing the number of groups in consensus clustering by pivotal methods. In: Perna, C., Salvati, N. Schirripa Spagnolo, F. (eds.) Book of Short Papers SIS 2021. ISBN 9788891927361 (2021)
14. Röblitz, S., Weber, M.: Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. Adv. Data Anal. Classif. **7**, 147–179 (2013)

15. Schiebinger, G., Wainwright, M.J., Yu, B.: The geometry of kernelized spectral clustering. Ann. Stat. **43**(2), 819–846 (2015)
16. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
17. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. Adv. Neural Inf. Process. Syst. **17** (2004)
18. Zhang, X., Li, J., Yu, H.: Local density adaptive similarity measurement for spectral clustering. Pattern Recognit. Lett. **32**(2), 352–358 (2011)

A Higher-Order PLS-SEM Approach to Evaluate Football Players' Performance

Mattia Cefis and Maurizio Carpita

Abstract Nowadays, data science is applied in several areas of our life, and also many applications in sport fields are increasing, following more and more a data-driven approach. In this context, we pay our attention on football (e.g. soccer in USA); by this project we aim to give a new approach and a statistical support in the evaluation of football players' performance provided from the EA Sports experts and available on Kaggle in the free KES dataset. For this purpose, we adopt a Higher-Order PLS-SEM approach by using *sofifa* and *fifacards* KPIs in order to compute a composite indicator and compare it with the well-known EA *overall* index provided by EA experts. Furthermore, we will compare two performance frameworks defined by experts, supporting them from a statistical point of view. The final goal is to give a statistical support and suggest a new performance indicator able at the same time to evaluate the overall performance and the key sub-areas (latent traits) of the players' performance. This is crucial for helping coaches and scouting staff of professional teams to take strategic decisions, in order to evaluate impartially players' performance and skills.

Keywords Football performance indicators · PLS-SEM · Composite indicators

1 Introduction

The latest developments in sports research, especially in football, are following a data-driven approach [6]. In order to give a clearer idea, in football analytics there are two main research approaches: an exploratory method oriented on analysis and classification of the KPIs (Key Performance Indices), that aims to evaluate players'

M. Cefis (✉) · M. Carpita
Department of Economics and Management, University of Brescia, Contrada S. Chiara, Brescia, Italy
e-mail: mattia.cefis@unibs.it

M. Carpita
e-mail: maurizio.carpita@unibs.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_4

performance [2, 4] and another one oriented on the prediction of football match results [5]. Furthermore, in order to evaluate player's performance there exist different methods: for example Pappalardo [18] adopted a SVM observing match outcome, Schultze and Wellbrock [21] created a rating performance index thanks to a plus-minus metric, Carpita et al. [3] adopted an unsupervised method to classify different areas of performance. We will focalize our attention on this last issue, in fact our goal is to explore players' KPIs, in order to evaluate some different strategic skills.

In particular, we know that players' performance evaluation is becoming a strategic key for football coaches and for the management of a team. We know that players' performance on the soccer field has been extensively measured and described by soccer experts: in literature, very important is the detailed classification by the experts from Electronic Arts (EA Sports)[1]. In their opinion, players' performance can be thought as a multidimensional construct made up of 6 performance composite indicators (*defending*, *attacking*, *mentality*, *movement*, *skill* and *power*), each one composed from several specific skills (e.g. *marking*, *standing tackle* and *sliding tackle* as elements for the *defending* dimension), which combined form the EA *overall* indicator that "sums up" the performance; here the main problem is that experts' opinion is not statistically supported [2, 4].

In this paper, our goal is to propose the use of an Higher Order approach in the Partial Least Squares Structural Equation Modeling (PLS-SEM), starting from the data provided by a relevant data science platform (Kaggle) in order to build a new composite indicator and to compare it with the well-known *overall* indicator provided by EA Sports experts, giving a significant statistics support to the experts' opinion. In particular, by PLS-SEM we aim to explore and compare two different experts' models of performance, named *sofifa* and *fifacards*.

2 The PLS-SEM and the Higher-Order Model

The PLS-SEM (Partial Least Squares Structural Equation Modelling [13, 23]) also known as PLS-PM (PLS Path-Modelling) is a non-parametric technique that belongs to a big family of statistical models useful to construct composite indicators: this topic had an exponential growing in the last 20 years, it is used to evaluate and supervise issues in a wide range of topics such as economy, society, industry and health [8]. In particular, a composite is formed when some indices are compiled into a single indicator on the basis of an underlying model, in fact it is used to measure multidimensional concepts; so, its purpose is to synthesise a phenomenon and monitoring it, in order to help policy makers to take strategic decisions [8].

[1] www.easports.com.

The core of the current technique was developed by Wold in 1982 [23], when he proposed his "soft-model basic design", underlying PLS-SEM as an alternative to the well-known covariance-based model [16]. Its goal is to measure causality relation between concepts (e.g. latent variables, the latent traits of the performance in our case), starting from some manifest variables (MVs, in our case the KPIs), thanks to an explorative approach: the explained variance of the endogenous latent variables (LVs, variables that we see as outcome, the performance in our case) is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression [17]. In addition, PLS-SEM does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. The PLS-SEM procedure estimates simultaneously two models:

- Measurement (or outer) model $\Rightarrow$ links MVs to their own LVs. Each block of MVs must contain at least one MV and this relation can be treated in two ways: reflective (where the MVs are the effects of their own LV) and formative (where the MVs are the causes of their own LV). In our work we will assume a formative structure for the outer model where each LV ξ_g is considered to be formed by its MVs following a multiple regression:

$$\xi_g = \mathbf{X}_g \mathbf{w}_g + \delta_g \tag{1}$$

  where $\mathbf{X_g}$ is the MVs matrix of the block g, and

$$E[\delta_g | \mathbf{X}_g] = \mathbf{0} \tag{2}$$

  where $\mathbf{w}_g$ is the vector of the outer regression weights and $\boldsymbol{\delta}_g$ is the vector of error terms. So, the vector of the outer weights for the g-th LV is estimated by least squares:

$$\mathbf{w}_g = (\mathbf{X}_g^T \mathbf{X}_g)^{-1} \mathbf{X}_g^T \xi_g \tag{3}$$

- Structural (or inner) model $\Rightarrow$ in this model LVs are divided into two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous. For the j-th endogenous variable in the model, the linear equation of its own structural model is:

$$\xi_j = \beta_0 + \sum_{r=1}^{R} \beta_{rj} \xi_r + \zeta_j \tag{4}$$

  where R is the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a linkage between the r-th exogenous LV and the jth endogenous LV and ζ_j is the error term.

Moreover, for our work we will assume a PLS-SEM with Higher-Order constructs, also known as hierarchical models [7, 9, 20]. In this framework we can include LVs that represent a "higher-order" of abstraction; since these LVs are virtuals, and so

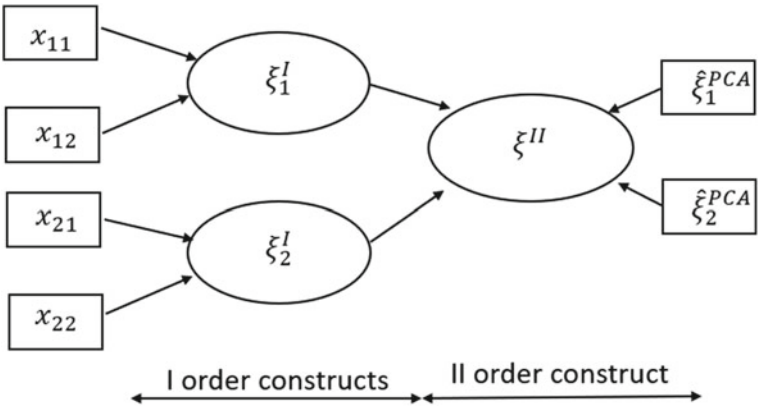


Fig. 1 The PLS-PM in the formative higher (second) order framework

without any apparent MVs, literature suggests us an interesting technique in order to manage this framework: a two-step or patch approach [20]. In the first step of this approach, we can compute by Principal Component Analysis (PCA) the scores of the lower-order LVs (the first principal component-I PC-of each one), while in the second one we can apply the classical PLS-SEM using the computed PCA scores as MVs for the endogenous LVs (Fig. 1). On the other hand, PLS-SEM shows some critical point, emphasize in different researches. For example Dijkstra and Henseler [10] proofed the inconsistency of PLS-PM in the reflective approach by showing adverse results in the hypothesis test. The solution proposed is a framework called PLS consistent (i.e. just for reflective constructs), that performs particularly well when the initial data are not normally distributed. Again, PLS-SEM is characterized by a lack of a global optimization procedure [15].

3 Data and Models Specification

In the European framework, the Kaggle European Soccer (KES) database is the biggest one devoted to the soccer leagues of European countries: it contains data about 10,000 players and 21,000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016. It is composed mainly by two big tables:

- The Match table contains the date, the positions (X and Y coordinates) on the pitch for the 22 players of the two teams and the final result of each match.
- The Player Attributes table contains other 29 variables (KPIs), with periodic player's performance on a 0–100 scale with respect to different abilities.

For this work, since it is an exploratory project, we took into account midfielder's players from Italian Serie A 2015/2016, with stats relying on the beginning of the sea-

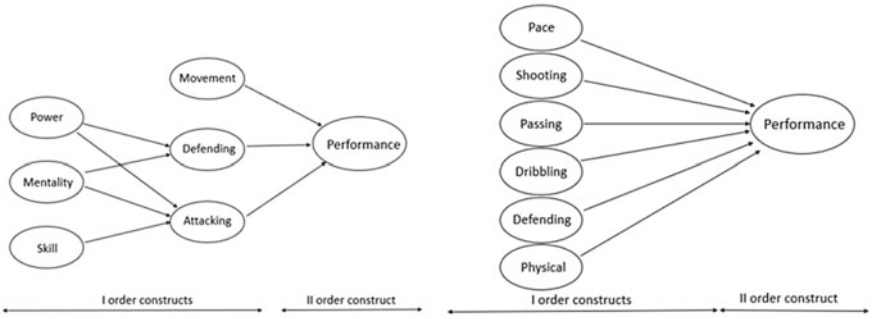


Fig. 2 PLS path *sofifa* model versus *fifacards* model

son: we obtained a dataset of 106 players and 29 KPIs for each one (see the Appendix for details). As said in the introduction, for what concern attributes' description, experts of EA Sports are considered the main authority: players' performance is defined as a multidimensional entity made up of 6 latent traits (e.g. *attacking*, *skill*, *movement*, *power*, *mentality*, *defending*), but they are not statistically supported [2, 4].

Taking in mind what we said in Sect. 2, for our purpose we assumed players' performance as extra-latent construct of higher (second) order. In particular we built two different models, following the experts' suggestion,[2] in order to replicate the EA Sports *overall*:

- In the first framework, with the classical *sofifa* LVs classification (6 groups of LVs), we assumed a conceptual structure behind the performance with the presence of 3 endogenous LVs, as suggested by some experts [11, 19, 20]: *attacking*, *defending*, and the player's performance (PLS path in Fig. 2). Note that for the performance (the only II order construct), we used the I PCs of *movement*, *defending* and *attacking* as MVs.
- In the second framework we took in consideration the *fifacards* classification (a little bit different classification of the same 29 MVs into others 6 LVs); here we assume just one endogenous Higher-Order LV (performance) influenced directly from the 6 exogenous (Fig. 2).

Sofifa and *fifacards* composite indicators have the same goal (exploring and measuring the football players' performance) but they follow two different experts philosophies in KPIs classification [1, 22]. For the application we used the R package *plspm* [20] and 1000 bootstrap resampling for the models validation. In the next section we will share our results.

[2] For details: www.fifauteam.com/fifa-19-attributes-guide.

Table 1 PLS-SEM higher-order goodness of fit

Model	GoF	Corr. with the EA *overall* index	95% CI for R^2 of the endogenous LVs
sofifa	0.71	0.94	$Attacking$: [0.89; 0.95] $Defending$: [0.67; 0.83] $Performance$: [0.95; 0.98]
fifacards	0.82	0.93	$Performance$: [0.98; 0.99]

4 Results and Discussion

We show results starting from the assessment indices by our two estimated PLS-SEM models, in order to understand their performance. Table 1 shows some quality indices for each endogenous LV and results of the bootstrap validation (a summary of the 1000 bootstrap samples stats, [20]): both models show a significant R^2 index for their own endogenous LVs. In the second model (i.e. *fifacards*) all the MVs and the LVs are significant, whereas for the first one there is just one non-significant MV and LV.

In Table 1 we can see also a comparison regards some assessments index between the two models: the goodness of fit index (GoF) is good (e.g. GoF > 0.7, [20]) and reveals that the second model is a little bit better than the first one. Then we computed the correlation coefficient r between our Higher-Order PLS-SEM performance indicator and the EA *overall*, and it shows a very high concordance ($r > 0.9$). All frameworks are considerable, supporting from a statistical point of view experts' opinion: following this, now we can deep our analysis considering the outer and the inner model for both models.

In order to study the outer models, we take in consideration the loadings factor for each MV in the two models. Respectively in Figs. 3a, b and 4a, we can see loadings and its 95% confidence interval (CI) after 1000 bootstrap resampling. We must take in mind that CIs including zero show non-significance for that specific MV: in particular, for the *sofifa* model the only non-significant variable is *Mov5*, whereas there aren't anyone for what concern the *fifacards* model.

Figure 3a reveals that the only LV poorly related with its own MVs is *movement* (4 KPIs with loading < 0.25, [20]): in fact it is strongly represented just from *Mov4*. The remaining MVs have significant and positive effects on its own (loading > 0.25) latent trait. Instead, regarding the *fifacards* model loadings (Fig. 4a, b), all MVs have significant and positive estimated coefficients. We can say that in both models the MVs are strongly related to the correspondent LV defined by experts, except *movement* for the *sofifa* framework.

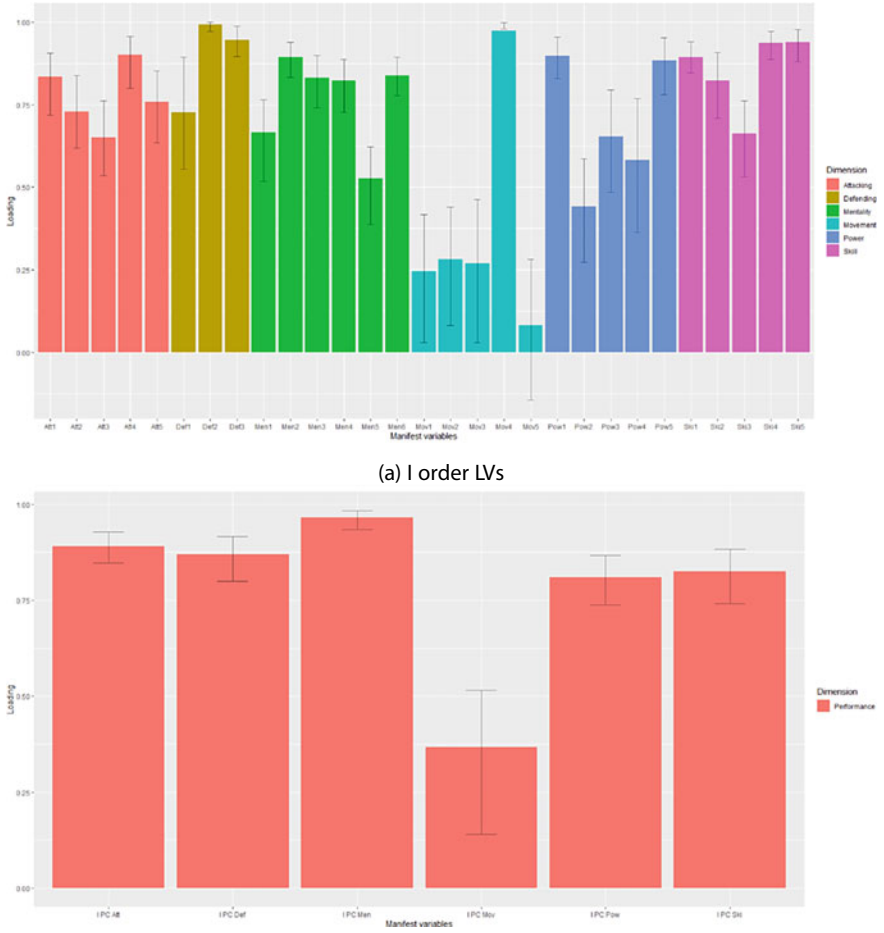
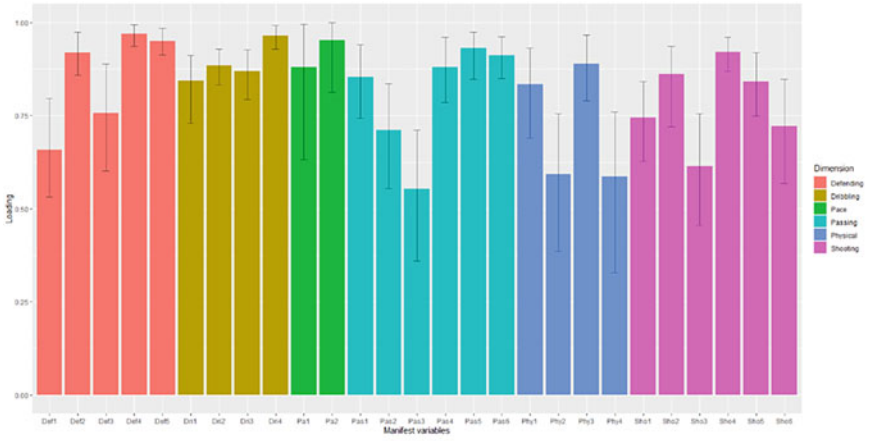
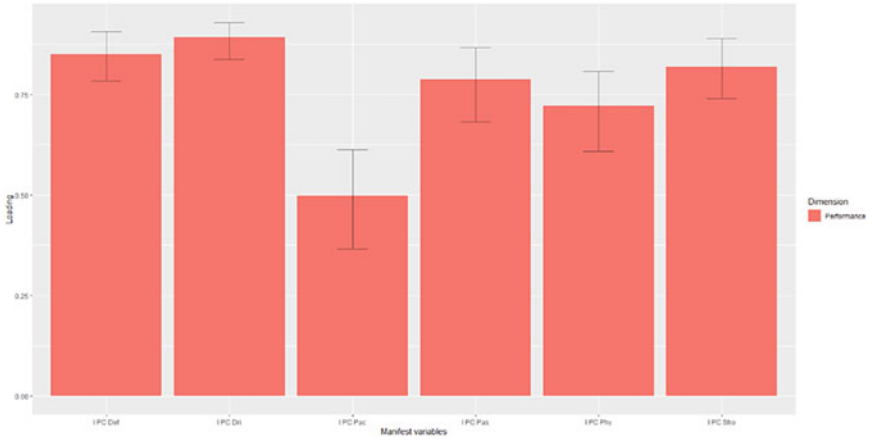


Fig. 3 I order LVs

At this point we focus attention on the inner model for each framework, by analysing their structure. In Fig. 5 we can see the *sofifa* inner model estimates obtained by our bootstrap validation; note that in red is indicated the only path non-significant: so, the *skill* latent trait is non-relevant from a statistical point of view for the *defending* ability of one midfielder, whereas it is the most important (in bold) for the *attacking*. For *defending*, the most important LV is *mentality* (0.76), while for the final aggregate indicator (i.e. *performance*) is *attacking* (0.50), following from *defending* (0.42) and outdistanced *movement* abilities (0.15).



(a) I order LVs



(b) II order LV

Fig. 4 I order LVs

Instead, in Fig. 6 is shown the path diagram for the *fifacards* model: all bootstrap path estimates are statistically significant. *Pace, shooting, passing* and *physical* have a similar impact on the composite *performance* indicator (between 0.15 and 0.23), whereas *dribbling* and *defending* have the strongest one (respectively 0.26 and 0.28).

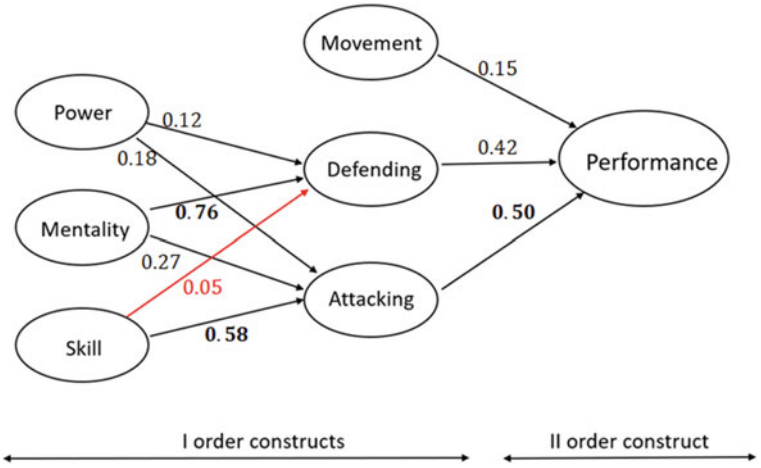


Fig. 5 The *sofifa* inner model path estimates

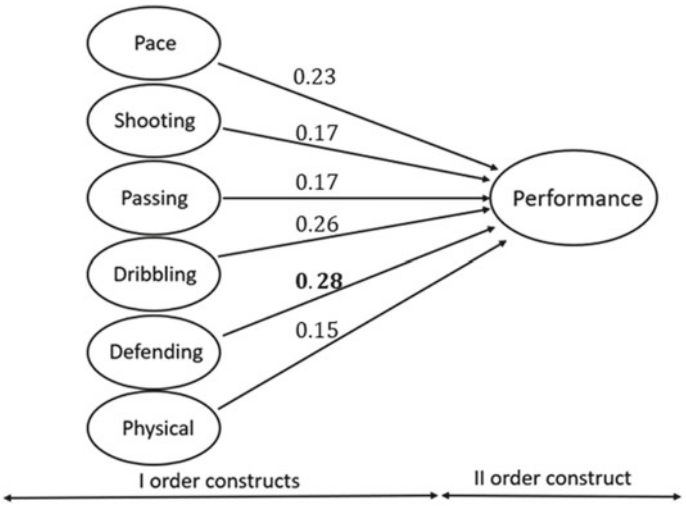


Fig. 6 The *fifacards* inner model path estimates

5 Conclusion

In this research we tried to explore and compare two interesting football performance frameworks provided by EA experts thanks a PLS-SEM second-order hierarchical model applied on a sample of 106 midfielders from Italian Serie A. Due to the lack

of statistical validation for what concern the experts' theory, our goal was to support them by this innovative approach in the field of football analytics application. We validated results by 1000 bootstrap resampling and we summarized the models performance by some assessing indices. As output, both models reached the goal to explore and measuring midfielders' performance, since they showed interesting and encouraging results (i.e. high correlation with the existing EA *overall* indicator) and medium-high GoF index (0.82 for *fifacards* and 0.71 for *sofifa*). In order to make an in-depth analysis, we verified the consistency and the significance about the outer and the inner models: for what concern the first one, all MVs had a strong and a statistically significant relation (loadings > 0.25) with their own LVs, except *mo5* for the *sofifa* model; about the *sofifa* inner model, the only one non-significant LV is *skill*. Again, analysing the midfielders' *sofifa* structural model we could emphasize how for *defending* the most important (i.e. with highest path coefficient) LV was *mentality*, while for the *attacking* phase was *skill*; finally, for the overall *performance* the weight was quite balanced between attacking and defending. Instead, regarding the midfielders' *fifacards* model, the overall *performance* was mainly influenced by *defending* and *dribbling*. These models could be useful for understanding any key choice of coaches, as well as to guide player transfer decisions, transfer fees or to improve future predictive models. We can consider this work as an explanatory starting point: first, it could be interesting for future projects to compare restults with other approaches, that outperform the lacks of PLS-SEM, maybe using Generalized Structured Component Analysis (GSCA, [15]) or more recent approaches to compute the higher order constructs, like the mixed two-step approach [9]. In addition, these models could taking in consideration heterogeneity observed across others European leagues and players' roles. Furthermore, a confirmatory tetrad analysis (CTA, [12]) for analysing the nature of the constructs (i.e. reflective or formative) and a confirmatory composite analysis (CCA, [14]) could be very interesting, with the final goal to introduce this new composite indicator in a predictive model.

Appendix

Here we show a statistical summary about the 29 abilities (MVs or KPIs) of EA Sports and their corresponded *sofifa* and *fifacards* classification (Appendix Table 2).

Table 2 The EA Sports KPIs statistics with *sofifa* and *fifacards* experts classification

EA KPIs (MV)	*sofifa* (LV)	*fifacards* (LV)	Mean	Std	Skew	Q1	Q2	Q3
Crossing	Attacking1	Passing1	64.85	9.19	−0.90	60	67	71
Finishing	Attacking2	Shooting1	56.72	10.53	−0.26	50	58	64
Heading accuracy	Attacking3	Defending1	58.52	10.01	−0.33	50	60	67
Short passing	Attacking4	Passing2	76.20	5.75	−0.64	74	77	80
Volleys	Attacking5	Shooting2	59.34	11.31	−0.37	51	60	68
Dribbling	Skill1	Dribbling1	72.58	6.85	−0.45	68	73	77
Curve	Skill2	Passing3	64.13	10.26	−0.46	57	65	72
FK accuracy	Skill3	Passing4	58.45	11.59	0.04	49	58	68
Long passing	Skill4	Passing5	72.86	5.31	−0.16	69	74	76
Ball control	Skill5	Dribbling2	75.08	6.40	−0.73	71	76	78
Acceleration	Movement1	Pace1	68.73	8.54	−0.99	66	69	75
Sprint speed	Movement2	Pace2	67.00	8.52	−0.73	65	68	71
Agility	Movement3	Dribbling3	71.00	7.67	−0.16	66	72	76
Reaction	Movement4	Dribbling4	70.38	7.64	−0.31	66	71	75
Balance	Movement5	Dribbling5	70.22	10.18	−1.05	65	70	78
Shot power	Power1	Shooting3	70.67	8.77	−0.51	64	72	77
Jumping	Power2	Physic1	64.75	10.82	−0.16	58	65	72
Stamina	Power3	Physic2	73.39	10.46	−0.40	67	75	79
Strength	Power4	Physic3	66.77	11.32	−1.01	64	68	75
Long shot	Power5	Shooting4	67.04	11.11	−1.17	63	69	74
Aggression	Mentality1	Physic4	67.67	11.23	−0.57	60	69	75
Interception	Mentality2	Defending2	66.08	11.53	−1.30	60	69	74
Positioning	Mentality3	Shooting5	66.73	8.95	−0.70	62	67	74
Vision	Mentality4	Passing6	71.24	6.96	−0.36	66	72	76
Penalties	Mentality5	Shooting6	57.59	9.56	0.00	51	57	65
Composure	Mentality6	Dribbling6	70.44	7.25	−0.30	66	71	75
Marking	Defending1	Defending3	64.19	10.92	−1.15	60	65	71
Standing tackle	Defending2	Defending4	66.23	10.72	−1.16	60	68	74
Sliding tackle	Defending3	Defending5	62.76	10.57	−0.67	58	63	70

References

1. Biecek, P., Burzykowski, T.: Explanatory Model Analysis: Explore, Explain and Examine Predictive Models. Chapman and Hall/CRC (2021)
2. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the Kaggle European soccer database. Stat. Model. **19**(1), 74–101 (2019)
3. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: a statistical perspective. Soc. Indic. Res. **156**(2), 815–830 (2021)
4. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the European soccer leagues. J. Appl. Stat. **48**(9), 1696–1711 (2021)
5. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the drivers of football match outcomes with data mining. Qual. Technol. Quant. Manage. **12**(4), 561–577 (2015)

6. Cefis, M.: Football analytics: a bibliometric study about the last decade contributions. Electron. J. Appl. Stat. Anal. **15**(1), 232–248 (2022)
7. Ciavolino, E., Carpita, M., Nitti, M.: High-order pls path model with qualitative external information. Qual. Quant. **49**(4), 1609–1620 (2015)
8. Commission, J.R.C.E., et al.: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing (2008)
9. Crocetta, C., Antonucci, L., Cataldo, R., Galasso, R., Grassia, M.G., Lauro, C.N., Marino, M.: Higher-order pls-pm approach for different types of constructs. Soc. Indic. Res. **154**(2), 725–754 (2021)
10. Dijkstra, T.K., Henseler, J.: Consistent partial least squares path modeling. MIS Quart. **39**(2), 297–316 (2015)
11. Filetti, C., Ruscello, B., D'Ottavio, S., Fanelli, V.: A study of relationships among technical, tactical, physical parameters and final outcomes in elite soccer matches as analyzed by a semiautomatic video tracking system. Percept. Motor Skills **124**(3), 601–620 (2017)
12. Gudergan, S.P., Ringle, C.M., Wende, S., Will, A.: Confirmatory tetrad analysis in pls path modeling. J. Bus. Res. **61**(12), 1238–1249 (2008)
13. Hair, J.F., Ringle, C.M., Sarstedt, M.: Pls-sem: indeed a silver bullet. J. Mark. Theory Pract. **19**(2), 139–152 (2011)
14. Hair, J.F., Jr., Howard, M.C., Nitzl, C.: Assessing measurement model quality in pls-sem using confirmatory composite analysis. J. Bus. Res. **109**, 101–110 (2020)
15. Hwang, H., Takane, Y.: Generalized structured component analysis. Psychometrika **69**(1), 81–99 (2004)
16. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika **43**(4), 443–477 (1978)
17. Monecke, A., Leisch, F.: Sempls: structural equation modeling using partial least squares. J. Stat. Softw. **48**(1), 1–32 (2012)
18. Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., Giannotti, F.: Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Trans. Intell. Syst. Technol. (TIST) **10**(5), 1–27 (2019)
19. Riboli, A., Semeria, M., Coratella, G., Esposito, F.: Effect of formation, ball in play and ball possession on peak demands in elite soccer. Biol. Sport **38**(2), 195 (2021)
20. Sanchez, G.: Pls Path Modeling with R. Berkeley. Trowchez Editions **383** (2013)
21. Schultze, S.R., Wellbrock, C.M.: A weighted plus/minus metric for individual soccer player performance. J. Sports Anal. **4**(2), 121–131 (2018)
22. Sherif, S.: Ea explains how Fifa player ratings are calculated. VG247. Com. **27** (2016)
23. Wold, H.: Encyclopedia of Statistical Sciences. Partial Least Squares, pp. 581–591. Wiley, New York (1985)

A Latent Markov Approach for Clustering Contracting Authorities over Time Using Public Procurement Red Flags

Simone Del Sarto, Paolo Coppola, and Matteo Troìa

Abstract Public procurement is a field particularly prone to corrupt acts. In this regard, red flag indicators are developed with the purpose of signalling anomalies and alerting the system for a potential risk of corruption. By exploiting data coming from the Italian National Database of Public Contracts, a set of red flag indicators for the measurement of the risk of corruption in public procurement are computed for a sample of contracting authorities and monitored over time. Afterwards, a latent Markov model for continuous responses is applied to the data at issue, in order to: i. identify groups of contracting administrations (i.e., the model latent states), characterised by different behaviours on the basis of the selected red flags, and ii. estimate the transitions over time across the ascertained groups. Results show that seven profiles of administrations may be highlighted on account of the selected red flags. Among them, four profiles include contracting authorities with indicator values associable to a low risk, whereas one profile can be labelled as the most at-risk one according to the respective indicator means.

Keywords Red flags · Corruption risk · Public procurement · Latent Markov

S. Del Sarto (✉)
Department of Political Science, University of Perugia, Perugia, Italy
e-mail: simone.delsarto@unipg.it

P. Coppola
Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
e-mail: paolo.coppola@uniud.it

M. Troìa
Capgemini Italia – Insights & Data, Milan, Italy
e-mail: matteo.troia@capgemini.com

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_5

1 Introduction

The exponential increase in computing power that we have experienced over the last sixty years has profoundly transformed the society in which we live. Digital transformation has completely changed and will continue to change so many aspects of human life and activity. Such a transformation prompts the question of whether and how we can use new technologies to more effectively and efficiently achieve our goals.

The range of activities deployed to address corruption is no exception. In the age of data, the availability of huge structured databases, automatically fed by increasingly powerful and high-performance information systems, makes it possible to refine the techniques of study and analysis not only with regard to the ex-post identification of misconduct, but also to carry out constant monitoring that can help prevent fraud and inefficiency.

Corruption, unfortunately, is a phenomenon that is very difficult to measure and detect, due to its inherent complexity [3, 13]. The parties who are harmed by corruption (the public authorities, the State and the community as a whole) do not participate in the corrupting event, hence they have no immediate perception of it. In retrospect it is possible to highlight cost increases or discover poor quality supplies, but these inefficiencies are far from being able to prove that corruption took place. In the act of corruption, the corrupter and the corrupted both have an interest in maintaining secrecy, because they both benefit from what they are doing. The best known index in this field is the Transparency International's Corruption Perception Index (CPI) [22], which is calculated on the perception of experts and business executives who are interviewed in 180 different countries and territories. The CPI is undoubtedly useful in terms of communication and raising awareness among the general public about the plague of corruption, but we know that the methodology based on interviews is affected by the cultural context and can be influenced by scandals and news spread by the media [11].

In order to be able to estimate corruption in a more objective manner in a given territory, the analysis of denunciations and convictions of such crimes can certainly be useful. However, this method suffers, on the one hand, from an underestimation due to the fact that, as already said, corruption is a particularly complex phenomenon to bring to light, and, on the other hand, the timing of justice can make it very complicated to have a stable measure of corruption that is sufficiently close in time. Above all, indices calculated on these data can only photograph a phenomenon that has already occurred and have limited utility in prevention. Similarly, the indices that attempt to estimate the adequacy of the expense sustained to carry out the works with the quality of the same [14] are difficult to apply, since the correct evaluation of the value of goods and services is not at all immediate and furthermore, they can also estimate corruption only after it has been perpetrated.

Anyway, recent approaches in corruption measurement rely on red flag indicators [2, 17], which signal risk of corruption, rather than actual corruption, and are supposed to be correlated with corrupt practices rather than perfectly matching

them [16]. This approach is based on the preventative dimension (rather than repressive) of corruption control. In fact, the purpose of corruption prevention is to detect potential weaknesses of a public organisation, with the aim of alerting the system for possible vulnerabilities and opportunities for malpractices [6, 12].

The red flag approach is particularly suitable in the public procurement (PP) field, where the wide availability of data relating to calls for tenders and public contracts allows for the construction of red flag indicators that can highlight anomalies during the process of tendering and execution of the construction or supply of goods and services [10]. The constant monitoring of these indicators—made possible by the automation of the collection of quality data deriving from the correct and complete digitisation of the procedures relating to public tenders—can itself represent a disincentive to the implementation of corrupt behaviour. The analysis can cover, for example, the timing of calls for tenders and the procedure types chosen, the total number or absence of participants, bidding discounts or variations during the course of the work, and thus can cover the entire period of time in which the phenomenon of corruption is likely to manifest itself. Obviously, anomalies are not necessarily linked to corruption, and inefficiencies could arise from organisational shortcomings or lack of expertise, but here too the methodology based on red flags is useful in drawing attention to the need to correct any errors as soon as possible.

Italy, which is the study context of this work, has identified the National Database of Public Contracts (henceforth BDNCP, which stands for *Banca Dati Nazionale dei Contratti Pubblici*) as a database of national interest and has established in article 62-bis of Legislative Decree no. 82/2005 that "in order to promote the reduction of administrative burdens deriving from information obligations and ensure the effectiveness, transparency and real-time control of administrative action for the allocation of public expenditure in construction, services and supplies, also with a view to respecting legality and the correct action of the public administration and preventing corruption, the National Database of Public Contracts managed by the National Anticorruption Authority is used". Such a determination by the Italian Parliament and Government provides us with a single database in which to find all the information relating to all the phases of the public procurement process, from the publication of the call for tenders to the awarding of contracts, to possible variations during the course of the work, up to the testing of the works, for all the public contracts stipulated in the country.

By exploiting data extracted from the BDNCP we are going to construct a set of red flags indicators [9, 23] with the aim of defining different profiles for contracting stations, in order to then analyse the evolution over time in terms of transition between one profile and another. Our proposal is to use latent Markov models [4, 24] for continuous responses, which allows us to concurrently reach the above two-fold purpose by: i. identifying groups (i.e., latent states) of contracting administrations, characterised by different behaviours in terms of red flag indicators (hence, different levels of corruption risk), and ii. estimating the transition probabilities across states.

This paper is organised as follows: Sect. 2 describes the database at issue and the statistical model employed for this work, Sect. 3 reports the main results and some concluding remarks are drawn in Sect. 4.

2 Data and Statistical Model

This section deals with the materials and methods. In particular, Sect. 2.1 illustrates the BDNCP and defines the red flag indicators considered for this work, whereas Sect. 2.2 introduces the latent Markov model for continuous data.

2.1 BDNCP and Red Flag Indicators in PP

For this work we use data from BDNCP, which, as stated above, contains information about the entire life cycle of every public contract that have involved Italian contracting bodies. BDNCP is managed by the Italian National Anticorruption Authority (ANAC) and is set up by the legislative decree no. 235 of 2010. The birth of this database marks a fundamentally important stage in the evolution of the tasks incumbent on the authorities and institutions responsible for monitoring and controlling the flow of public money, since it represents an awareness by Italian institutions about the value and potential of data.

The systematic collection of every step of the complex process by which every single public contract is developed represents the first fundamental step in making the control of financial flows by the competent authorities timely and effective. In fact, BDNCP represents a single container of data from public bodies that are geographically distant and different in terms of size, tasks, areas and competencies. In this sense, this database can be considered a powerful tool for decision making, taking also into consideration that the data are publicly accessible and downloadable through the ANAC open data platform.[1]

Specifically, in this work we consider only procedures above 40,000 Euro, since, according to Italian law, most of the information for building red flag indicators in PP are recorded only for these kind of procedures (labelled as "ordinary" contracts).

Among the indicators already proposed in literature, five red flags are selected and then calculated for each contracting authority, in order to deal with at least one indicator for each of the main phases of the PP process (publication, bidding, award), hence considering at least one typology of risk for each step. The selected indicators are the following:

1. *non_open*: proportion of non-open procedures. The related information can be retrieved from the procedure type contained in the call for tender. In particular, according to this field, each procedure can be classified into open, restricted, negotiated or direct award. Accordingly, the last two categories are recognised as non-open. A high value of this indicator signals a risk factor as competition is limited. On the other hand, there may be a connection with the financial size of the tenders, and it may be insignificant in the case of contracting authorities that have issued an extremely low number of tenders in the year;

[1] https://dati.anticorruzione.it/.

2. *single_bid*: proportion of procedures for which a single bid is received. Using the award notice data, a single bid procedure is a procedure for which a unique bid has been received and declared as feasible by the contracting authority. This indicator, when values are high, also represents an anomaly in the functioning of the market, such as the presence of oligopolies or cartels, which could indicate a risk of corruption;
3. *non_price_eval*: proportion of procedures awarded through non-price related evaluation criteria (such as the "economically most advantageous offer" criterion), hence with a certain degree of subjectivity. The information at issue can be retrieved from the award notice data, in particular in the field related to the award criterion. In this case, the risk of corruption is proportional to the degree of subjectivity, although high values of this indicator may indicate a particular attention to the quality of the supply by the contracting authority;
4. *publ_deadline*: average number of days between the publication of the call for tender and the bid submission deadline. Both dates are included in the call for tender data. An extremely limited number of days available for bidding creates a greater barrier to entry for new providers, again limiting competition and therefore can be considered a risk factor. On the other hand, there may be a relationship with the financial magnitude and complexity of the work or services required;
5. *award_deadline*: average number of days between the award notice (contained in the award notice data) and the bid submission deadline. In this latter case, the risk factor lies in an unwise assessment of bids that could be indicated by an extremely low number of days. This indicator may also be affected by the type of tender, because, for example, tenders with maximum discounts can be evaluated more quickly than those with the criterion of the economically most advantageous offer. Other factors (like the previous indicator) may also be the complexity of the work or services required, as well as the financial size. Of course, there may also be a connection with the number of bids received, the efficiency of the contractor or with litigation, two aspects which, however, are not present in the database.

Indicators 1, 2 and 3 are such that the greater the indicator, the higher the risk can be considered, whereas the opposite occurs as regards indicators 4 and 5. Finally, data about public contracts with a call published from 2015 to 2017 are considered in this work: the above indicators are then computed for each contracting authority and each year (subject to data availability). More details on the indicator computation and evaluation are given in Sect. 3.

2.2 The Latent Markov Model

The latent Markov (LM) model is one the most widely known latent variable model. It is very useful in analysing longitudinal data, resulting from those situations in which the statistical units are followed over time according to one or more outcomes. LM models allow for the dynamic clustering of the statistical units into groups, also

called latent states, characterised by average values of the outcome(s). Moreover, these models are able to estimate the transitions across states over time, that is, given the "starting" group, the probabilities of moving towards other groups. These transitions may be constant over the follow-up period (time-homogeneous transitions) or different for each follow-up time period (time-heterogeneous transitions). For example, if units are observed over five years, transition probabilities across states may be different from year to year, or, conversely, a unique set of transition probabilities can be estimated, which is therefore common over years.

LM models are generally suitable for modelling categorical outcomes, for which latent states are characterised by conditional (to the state) response probabilities (for each category minus one). LM models for continuous variables are also allowed, thus latent states may be described in terms of outcome means. In our case, as the red flag indicators introduced in Sect. 2.1 may considered as continuous variables, the version of LM model that considers multivariate continuous outcomes is presented in the following of this section. Moreover, the statistical units are the contracting authorities that publish the call for tenders, while the latent phenomenon used for their clustering over time is the corruption risk in PP (of which red flag indicators are manifest variables).

Let Y_{ijt} be outcome j related to statistical unit i at time t, with $j = 1, \ldots, J$ (the red flags, in our case), $i = 1, \ldots, n$ (contracting authorities) and $t = 1, \ldots, T$ (years). Outcomes may be arranged in a J-dimensional response vector $\boldsymbol{Y}_{it} = [Y_{i1t}, \ldots, Y_{iJt}]^\top$ for each statistical unit. Furthermore, we suppose that a latent process affects response $\boldsymbol{Y}_{it}$, denoted with $\boldsymbol{V}_i = [V_{i1}, \ldots, V_{iT}]^\top$ and is assumed to follow a first-order Markov chain with k latent states. Hence, given latent process $\boldsymbol{V}_i$, the response vector is assumed to be normally distributed, with state-specific mean vector, denoted with $\boldsymbol{\mu}_v$, and common (across states) variance-covariance matrix, Σ:

$$f(\boldsymbol{Y}_{it} = \boldsymbol{y} | V_{it} = v) \sim N(\boldsymbol{\mu}_v, \Sigma),$$

where $\boldsymbol{y}$ and v are realisations of $\boldsymbol{Y}_{it}$ and V_{it}, respectively.

The parameters of a LM model for multivariate continuous variables are the following. For the measurement model we have the conditionals means, $\boldsymbol{\mu}_v$, with $v = 1, \ldots, k$ and the variance-covariance matrix, Σ. The latent model, on the other hand, has the following set of parameters:

- the initial probabilities, $\pi_v = P(V_{i1} = v)$, with $v = 1, \ldots, k$: they represent the probability of being in each state at the beginning of the study (i.e., when $t = 1$);
- the transition probabilities, $\pi^{(t)}_{v|\tilde{v}} = P(V_{it} = v | V_{it-1} = \tilde{v})$, with $t = 2, \ldots, T$ and $v, \tilde{v} = 1, \ldots, k$: they are generally arranged in a matrix and consist in the probability of moving towards state v given that in the previous time period we were in state $\tilde{v}$. In this case, we consider time-heterogeneous transition probabilities, hence different for each time period $t = 2, \ldots, T$.

Given the above, the distribution of the latent process has the following probability mass function:

$$P(\boldsymbol{V}_i = \boldsymbol{v}_i) = \pi_{v_{i1}} \prod_{t=2}^{T} \pi^{(t)}_{v_{it}|v_{it-1}},$$

where $\boldsymbol{v}_i = [v_{i1}, \ldots, v_{iT}]^\top$ is a realisation of $\boldsymbol{V}_i$. Moreover, the conditional distribution of $\boldsymbol{Y}_i$ given the latent process is

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{V}_i = \boldsymbol{v}_i) = \prod_{t=1}^{T} f(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} | V_{it} = v_{it}),$$

with $\boldsymbol{y}_i = [\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}]^\top$, where each single realisation vector $\boldsymbol{y}_{it}$ has been already defined in terms of the corresponding random vector $\boldsymbol{Y}_{it}$.

Finally, the manifest distribution of $\boldsymbol{Y}_i$ is given by

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \sum_{\boldsymbol{v}_i} P(\boldsymbol{V}_i = \boldsymbol{v}_i) P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{V}_i = \boldsymbol{v}_i).$$

A maximum likelihood approach through an Expectation-Maximisation (EM) algorithm [8] may be used for estimating the model parameters. However, log-likelihood function multimodality issues may arise, then we perform a combination of deterministic and random initialisations of model parameters. Details on the EM algorithm may be found in [4]. Finally, the R package "LMest" [5] is used for fitting the model to the available data.

3 Results

Data from BDNCP about public contracts published in 2015, 2016 and 2017 are freely downloaded from the ANAC open data platform. Information about the call for tender are binded together and then merged with the award notice data using the tender unique identifier (CIG, *Codice Identificativo Gara*). Overall, we initially deal with about 300,000 procurement contracts and more than 12,000 contracting authorities. However, not all these contracts are suitable for the computation of the selected indicators. In this regard, we can observe Table 1, where, for each indicator, the number of procedures suitable for its computation is reported, together with the number of involved contracting authorities.

As we can see, almost all the procedures are included in computing indicator *non_open*, as useful data can be found in the call for tender, thus present for almost every contract. Conversely, when computing indicator *single_bid*, less than 20% of the entire dataset is exploited. In fact, several conditions must arise when building this indicator, such as the presence of award notice, the number of offering companies and the number of feasible bids for each awarded procedure. Moreover, this indicator needs to be computed only for open procedures and after having removed those for

Table 1 Number of procedures and involved contracting authorities for each indicator: the percentage is computed with respect to the initial amount of procedures in the dataset (299,376)

Indicator	Procedures		Contracting authorities
	N	%	
non_open	298,778	99.8	12,263
single_bid	59,083	19.7	6,070
non_price_eval	233,921	78.2	11,518
publ_deadline	72,710	24.3	11,131
award_deadline	70,905	23.7	10,945

which a unique bidder is requested by law (i.e., directs awards, identified using the procedure type field). Accordingly, indicator *non_price_eval* can be evaluated only for the contracts with an award notice and, within these, only for those having a value present in the evaluation criterion field. Furthermore, according to the ANAC working protocol, only the following ways of realisation of the procedure need to be considered (using the field with the same name): tenders, acquisitions in economy, work concession contracts, framework agreements and conventions. Moreover, direct awards need to be excluded when computing these indicators. Finally, indicators *publ_deadline* and *award_deadline* need to be computed only for open procedures. Moreover, after having removed the procedures with negative differences between the involved dates (less than 2% of the procedures), useful contracts for these indicators are around a quarter of the initial dataset.

The LM analysis is performed by considering only those contracting bodies for which the five selected indicators are available in all the three years (hence, no missing value is allowed), for a total of 1,202 contracting authorities. Figure 1 reports the boxplots describing the distribution of each indicator in each year. As we can see, the distribution of indicator *non_open* remains quite constant over years, as well as those of indicators *publ_deadline* and *award_deadline*. Conversely, an evident shift towards higher values arises for indicators *single_bid* and *non_price_eval*. Furthermore, as expected, the distribution of indicator *publ_deadline* appears more shrunk towards smaller values than those for indicator *award_deadline*, because of the award process that generally requires more time, as it involves the evaluation of received bids by contracting bodies.

The LM model for continuous data, introduced in Sect. 2.2, is then fitted to the data at hand. A first, broadly-known, issue about this modelling framework concerns the selection of the number of latent states (k), which is tackled by means of a mix of objective and subjective criteria. As far as the former is concerned, we rely on standard penalised likelihood criteria, such as the Bayesian Information Criterion (BIC) [19] and the Akaike Information Criterion (AIC) [1]. Specifically, several LM models are fitted with increasing number of latent states (from 1 to 12 in this case): as usually, the best model would be the one with the smallest information criterion. Table 2 shows the BIC and AIC values for each of the 12 estimated models (with k

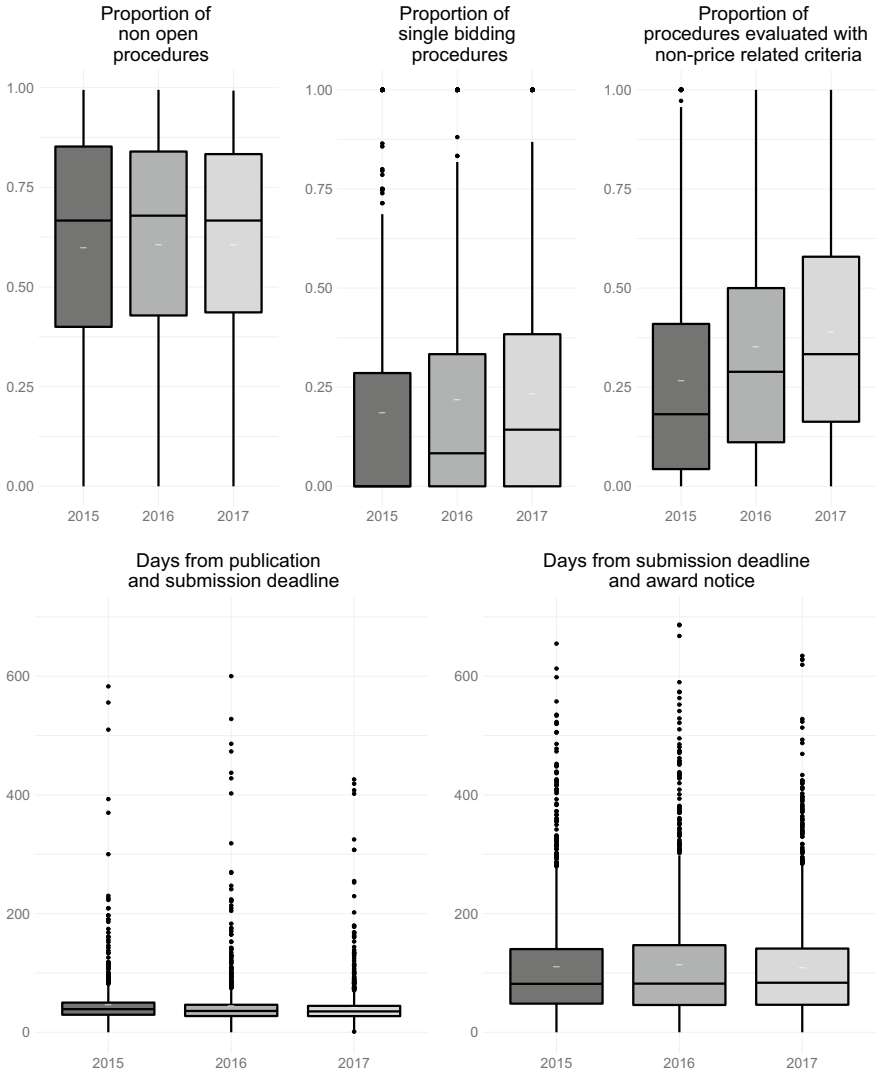


Fig. 1 Boxplots of each indicator by year (white dash represents the mean)

from 1 to 12), together with the relative difference (%) of each model with respect to the previous one with one latent state less. As can be noticed, the use of BIC would lead us to select $k = 10$, whereas $k > 12$ would be the best solution according to the AIC. However, such a large number of latent states generally leads to a blurred interpretation of the resulting profiles. In this case subjective judgements help us to reach the final solution, which needs to be selected with particular attention, also by considering criteria aimed at a clear interpretation of the results obtained [7, 15, 18].

Table 2 Output of the selection procedure of the number of latent states (k): % BIC and % AIC report the relative difference of each model with respect to the previous one (with $k-1$ latent states)

k	BIC	AIC	% BIC	% AIC
1	84,880.5	84,778.7	–	–
2	83,460.7	83,308.0	−1.67	−1.73
3	80,646.9	80,422.8	−3.37	−3.46
4	79,778.2	79,462.5	−1.08	−1.19
5	78,720.9	78,293.2	−1.33	−1.47
6	78,335.7	77,775.6	−0.49	−0.66
7	77,054.9	76,342.0	−1.64	−1.84
8	76,679.5	75,793.5	−0.49	−0.72
9	76,519.6	75,440.1	−0.21	−0.47
10	76,500.1	75,206.8	−0.03	−0.31
11	76,505.7	74,978.2	0.01	−0.30
12	76,626.8	74,844.7	0.16	−0.18

In fact, it is possible that a too large number of latent states, even if supported by statistical indices that confirm the optimality, may not be the most plausible choice.

Therefore, looking again at Table 2 and using the so-called elbow method [20, 21], it can be noticed that, despite the optimal solution is $k = 10$ using the BIC (or $k > 12$ using the AIC), the relative gain in terms of model fitting—as measured by % BIC or % AIC—with the increase of k can be considered negligible from eight latent states on out, as the relative improvement of the model fitting is less than 1%, both using the BIC and the AIC (see also Fig. 2, showing the BIC values against the number of latent states).

Consequently, $k = 7$ latent states are finally selected, allowing us to highlight seven profiles of Italian contracting authorities. They can be described by considering Table 3, which reports, for $v = 1, \ldots, 7$, the estimated conditional means, $\hat{\boldsymbol{\mu}}_v$ (by column), whereas the last row contains initial probabilities $\hat{\pi}_v$. Conditional mean vectors for each latent state are also represented in Fig. 3 using seven radar charts, in which the dark grey polygon, depicted in the centre of each panel, connects the grand means (hence, it is the same in each one).

Looking at the estimated initial probabilities, reported in the last row of Table 3, in 2015 (first year of analysis) the most numerous groups are latent states 2 and 5: the former includes approximately 20% of administrations, while more than half of contracting bodies belongs to the latter state (58%). On the other hand, latent states 1 and 3 are the smallest ones, as less than 2% of the involved contracting authorities are clustered into them. The other three states (4, 6 and 7) almost equally divide up the remaining share.

As far as the characterisation of each latent state is concerned, the following comments can be drawn:

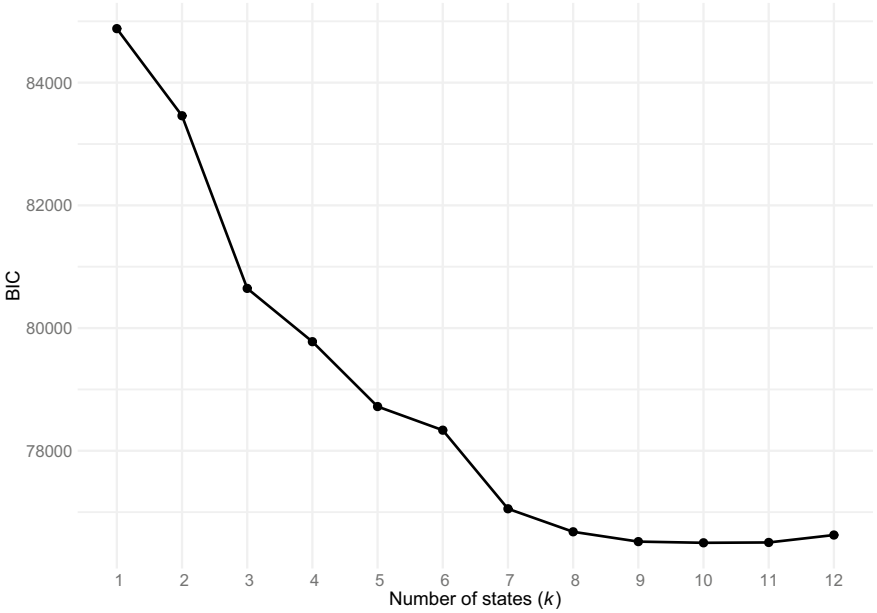


Fig. 2 BIC values against the number of latent states (k)

Table 3 Latent state descriptive parameters: estimated conditional means $\hat{\boldsymbol{\mu}}_v$ and initial probabilities $\hat{\pi}_v$, $v = 1, \ldots, 7$. Each indicator grand mean is obtained by averaging the means of each latent state and weighting by $\hat{\pi}_v$

Indicator	Latent state v							Indicator grand mean
	1	2	3	4	5	6	7	
non_open	0.609	0.246	0.697	0.633	0.740	0.299	0.655	0.600
single_bid	0.152	0.107	0.171	0.177	0.146	0.137	0.851	0.196
non_price_eval	0.274	0.143	0.332	0.476	0.278	0.771	0.385	0.307
publ_deadline	524.9	33.9	188.5	54.2	40.6	35.6	34.3	44.0
award_deadline	115.6	81.0	154.4	428.9	103.0	94.4	63.3	113.2
$\hat{\pi}_v$	0.004	0.193	0.019	0.053	0.580	0.071	0.079	

- latent state 1 (initial probability $\hat{\pi}_1 = 0.004$) includes contracting authorities whose first three indicators (related to the proportions of non-open contracts, single bids and procedures awarded with non-price related criteria) and the indicator *award_deadline* are, on average, around or below the grand means, whereas the opposite occurs for indicator *publ_deadline*. In particular, this group, although very few in number, presents the greatest value as regards the average number of

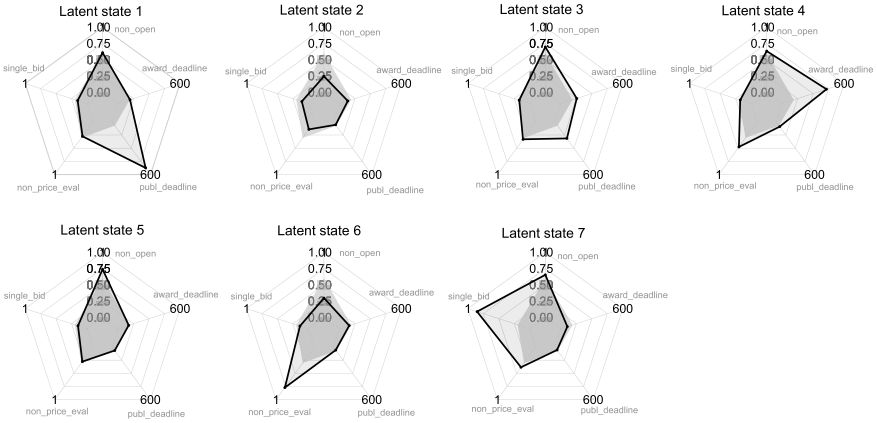


Fig. 3 Conditional means for each latent state, depicted through radar charts (dark grey polygons connect the grand means)

days between the publication of the call for tenders and the deadline for submitting a bid (about 525 days), very far from the indicator grand mean (44 days);

- latent state 2 (initial probability $\hat{\pi}_2 = 0.193$) reports the smallest values as regards the means of the first three indicators, but it has quite small average values also for the last two indicators. Hence, this is a group of administrations with very quick procedures, both in the publication and in the awarding phase, despite the fact that the proportions of non-open, single bid and non-price related tenders are around half the average value;
- latent state 3 (initial probability $\hat{\pi}_3 = 0.019$), like the first one, has a very long period (on average) between the publication of the call for tenders and bid submission deadline (188 days), but even the mean of indicator *award_deadline* (around 154 days) is far above the respective grand mean. The other three indicators are instead around the grand means (though *non_open* is rather above);
- latent state 4 (initial probability $\hat{\pi}_4 = 0.053$) has the greatest average distance between award notice and bid submission deadline (as measured by indicator *award_deadline*), equal to around 428 days, and is characterised by a consistent above-average use of non-price related criteria during the award phase (mean of *non_price_eval* equal to 0.476);
- the distinguishing feature of latent state 5 (initial probability $\hat{\pi}_5 = 0.580$) is related to a very frequent use of non-open procedures, as it present the greatest mean as regards indicator *non_open*, equal to 0.740 (even if not so far from the grand mean of 0.600);
- latent state 6 (initial probability $\hat{\pi}_6 = 0.071$) has the greatest mean of indicator *non_price_eval* (0.771), but it is one with the lowest mean as regards indicator *non_open*;
- latent state 7 (initial probability $\hat{\pi}_7 = 0.079$) is characterised by very low means of the indicators *publ_deadline* and *award_deadline* (hence, very quick procedures).

Also, it presents the highest (average) frequency of procedures with a single bid (0.851), very far from the other groups.

The transition matrices, which contain estimated transition probabilities $\hat{\pi}_{v|\tilde{v}}$, $v, \tilde{v} = 1, \ldots, 7$, are reported in Table 4 and depicted in Fig. 4 using directed graphs: the nodes represent the latent states, while the directed edges consider the transitions across states. Moreover, in the picture, only edges representing probabilities greater than 0.2 are reported (for clarity reasons) and the greater the edge thickness, the greater the underlying transition probability.

In particular, if we are in state 1, we can observe non negligible probabilities of moving towards states 3, 4 and 5 in the 2015–2016 transition (0.410, 0.203 and 0.206, respectively); however, the transition to state 4 turns out to be very unlikely during the 2016–2017, while those towards states 3 and 5 become more pronounced (equal to 0.499 and 0.502, respectively).

On the other hand, latent state 2 is a quite persistent state as $\hat{\pi}_{2|2} > 0.50$ in both time transitions. Besides, if we are in latent states 3 or 4, two alternatives are very frequent: to remain in the current state, or to move towards state 5. The latter state, together with the sixth, is highly persistent, as $\hat{\pi}_{5|5} > 0.75$ and $\hat{\pi}_{6|6} > 0.65$ in both time transitions. Finally, if we are in latent state 7, we can observe high probability to move towards state 5 (probabilities close to 0.50), although some chances to remain in the seventh state or to move towards the sixth one (related probabilities, however, not greater than 0.30) can be noticed.

Finally, we can observe a high overall tendency to move towards state 5. This is quite evident by looking at the fifth column of Table 4 (which reports $\hat{\pi}_{5|\tilde{v}}$, $\tilde{v} = 1, \ldots, 7$), or by observing node 5 in Fig. 4. In fact, we can notice that, almost for all the states, the probability to arrive at state 5 is very high. The only exceptions consist in latent states 2 and 6 (persistent states, as outlined above), together with state 4, though not so evident.

It is not straightforward to characterise the seven states in terms of risk. Starting with the one with the highest initial probability, that is, state 5—in which the absolute majority of contracting stations can be found—we can observe indicator *non_open* with the highest value compared to the other states. All other indicators have values around the grand mean. The risk is not negligible, but considering the size of the state and the high probability of persistence and absorption from states 1, 3, 4, and 7, it may be related more to a search by governments for simplified procedures that can reduce bureaucratic workload.

State 2—the one with the second highest initial probability—is the one with all indicators below the grand mean and is also characterised by a high probability of persistence. Although very low values of indicators *publ_deadline* and *award_deadline* might indicate a risk factor, in this case, although lower, they do not deviate much from the average and this state might be the one that represents efficient contracting stations, with low risk of corruption.

Scrolling through the latent states always according to the decreasing order of initial probability, state 7 is the one with which the highest risk can be associated, given the very high value of the proportion of bids with a single bidder that can only

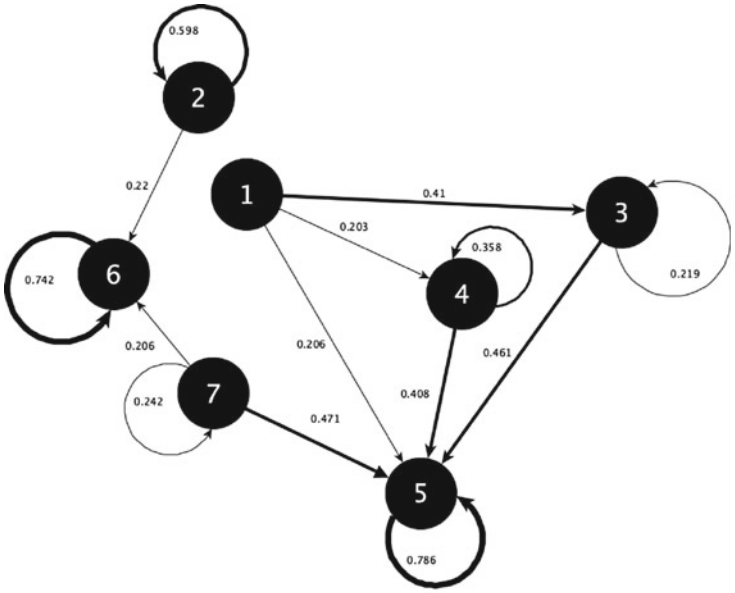
Table 4 Matrices of estimated transition probabilities ($\hat{\pi}_{v|\tilde{v}}$) from 2015 to 2016 (a) and from 2016 to 2017 (b). The "starting" states ($\tilde{v}$) are reported on the rows, while those of "arrival" (v) on the columns

(a) Transitions from 2015 to 2016							
	v						
$\tilde{v}$	1	2	3	4	5	6	7
1	0.000	0.000	0.410	0.203	0.206	0.000	0.181
2	0.008	0.598	0.015	0.023	0.081	0.220	0.055
3	0.133	0.050	0.219	0.137	0.461	0.000	0.000
4	0.000	0.000	0.027	0.358	0.408	0.085	0.123
5	0.004	0.024	0.017	0.048	0.786	0.014	0.107
6	0.000	0.125	0.000	0.028	0.000	0.742	0.104
7	0.000	0.010	0.033	0.038	0.471	0.206	0.242
(b) Transitions from 2016 to 2017							
	v						
$\tilde{v}$	1	2	3	4	5	6	7
1	0.000	0.000	0.499	0.000	0.502	0.000	0.000
2	0.006	0.642	0.004	0.021	0.153	0.108	0.067
3	0.000	0.000	0.374	0.000	0.556	0.038	0.033
4	0.014	0.015	0.028	0.385	0.380	0.073	0.105
5	0.003	0.000	0.008	0.030	0.840	0.027	0.092
6	0.000	0.119	0.008	0.010	0.065	0.671	0.127
7	0.000	0.000	0.007	0.032	0.447	0.205	0.309

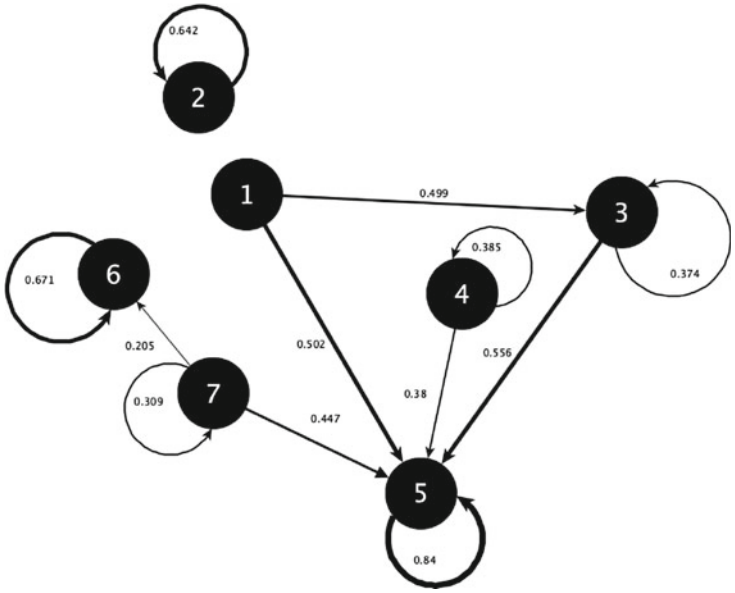
be partially explained by the above-average value of *non_open*. The probability of transition to states considered at lower risk is also particularly low.

State 6 also presents a non-negligible risk profile, due to the high value of the proportion of tenders with the award criterion not exclusively linked to price, combined to values of the publication and award deadlines below the grand mean and, like in the previous case, the probability of transition to states that can be considered at low risk is marginal.

Finally, states 4, 3, and 1 can be considered at low risk. State 4, with the particularly high adjudication deadline and a *non_price_eval* value above the grand mean, may represent contracting authorities that have had to sustain appeals. The transition probabilities are also consistent with this reading because the probability of persistence is not particularly high. State 3 and state 1 represent an extremely small number of contracting stations with very low persistence in transitions. They likely represent states with errors in the deadline data (especially state 1) or related to extraordinary events related to the formation of tender commissions.



(a) Transitions from 2015 to 2016.



(b) Transitions from 2016 to 2017.

Fig. 4 Graphical representation of the transition matrices reported in Table 4

4 Conclusions

This work aims at studying the behaviour of the Italian contracting authorities in terms of management of the public procurement process, phenomenon that can measured through specific red flag indicators. As such, this kind of indicators is able to measure the risk of corruption (rather the actual one) and consists in one of the most recent tools for fighting corruption, which focuses on its prevention instead of its repression.

Data on Italian public procurement about years 2015, 2016 and 2017 are extracted from the Italian National Database of Public Contracts, a huge database of national interest that contains detailed information about every single public contract regarding Italian public administrations. These data are then used for constructing some of the red flag indicators for measuring corruption risk in public procurement, according to the relevant literature on this topic. Specifically, the selected indicators are able to evaluate the corruption risk in each of the main steps of the public procurement process, and are the following: i. proportion of non-open procedures; ii. proportion of procedures with a unique bid; iii. proportion of procedures awarded using criteria not based on price; iv. average time interval (in days) between publication of the call for tender and bid submission deadline v. average time interval (in days) between award notice and bid submission deadline.

Using a latent Markov (LM) approach for multivariate continuous response variables, two objectives are concurrently reached, that is, i. the identification of groups of contracting authorities, characterised by specific mean values of the red flags at issue (hence, potential clusters with certain degrees of corruption risk in public procurement), and ii. the estimation of the transition probabilities over time across the ascertained groups.

Main findings pinpoint seven clusters of contracting bodies (i.e., the latent states of the LM model). In particular, indicator means of contracting authorities clustered into latent states 1, 2, 3 and 4 assume values that can reasonably be considered at low risk, or, anyway, around the sample grand mean. On the other hand, latent state 7 can be portrayed as the group of administrations having indicator average values associable to higher risk. However, looking at the estimated transition probabilities, latent state 1 is not a persistent state (probabilities to remain in this state very close to 0), while we estimate high probabilities to remain in the second group. What is more, among the ascertained clusters, latent state 5 results to be the one that is able to "grasp" most of the other states and is characterised by a high use of non-open procedures, whereas the other indicators lie around the average levels.

The present work is meant to be a first attempt to analyse the evolution over time of the behaviour of Italian public administrations in the public contract management. However, it presents some weaknesses, among them we can mention the dealing of missing values in the indicators (for certain statistical units and time periods), as well as the limited number of indicators (five) employed for describing the contracting authority behaviour.

Several considerations are going to be taken into account in order to strengthen the results. Other than addressing the drawbacks just outlined by extending the statistical

analysis accordingly, covariates describing peculiar features about the contracting bodies could be added in the modelling approach, in order to better characterise the latent states. Moreover, a quantile regression approach could be very useful in this context—therefore, an analysis that considers the tails of the indicator distributions—since the focus is to spot extreme behaviours, which manifest through high or low values of certain red flags.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, pp. 267–281. Akademinai Kiado (1973)
2. ANAC: Corruzione sommersa e corruzione emersa in Italia: modalitài di misurazione e prime evidenze empiriche. https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%20Assets/anacdocs/Attivita/Pubblicazioni/RapportiStudi/Metodologie-di-misurazione.pdf (2013). Accessed 4 Jan 2022
3. Andersson, S., Heywood, P.M.: The politics of perception: use and abuse of transparency international's approach to measuring corruption. Polit. Stud. **57**(4), 746–767 (2009)
4. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. CRC Press, Boca Raton, FL (2013)
5. Bartolucci, F., Pandolfi, S., Pennoni, F.: LMest: an R package for latent Markov models for longitudinal categorical data. J. Stat. Softw. **81**(4), 1–38 (2017)
6. Carloni, E.: Misurare la corruzione? Indicatori di corruzione e politiche di prevenzione. Politica del diritto **3**, 445–466 (2017)
7. Del Sarto, S.: L'utilizzo dei modelli IRT multidimensionali per la costruzione di profili di studenti. In Falzetti P (ed.): I risultati scolastici: alcune piste di approfondimento - III Seminario "I dati INVALSI: uno strumento per la ricerca", pp. 93–112. Franco Angeli, Milano (2021)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–22 (1977)
9. Fazekas, M., Tóth, I.J., King, L.P.: An objective corruption risk index using public procurement data. Eur. J. Crim. Policy Res. **22**, 369–397 (2016)
10. Fazekas, M., Cingolani, L., Tóth, B.: A Comprehensive Review of Objective Corruption Proxies in Public Procurement: Risky Actors, Transactions, and Vehicles of Rent Extraction. Government Transparency Institute Working Paper Series No. GTI-WP/2016:03, Budapest (2017)
11. Fiorino, N., Galli, E.: La corruzione in Italia. Il Mulino, Bologna (2013)
12. Gallego, J., Rivero G., Marínez, J.: Preventing rather than punishing: an early warning model of malfeasance in public procurement. Int. J. Forecast. **37**(1), 360–377 (2021)
13. Gnaldi, M., Del Sarto, S., Falcone, M., Troìa, M.: Measuring corruption. In: Carloni, E., Gnaldi, M. (eds.) Understanding and Fighting Corruption in Europe—From Repression to Prevention, pp. 43–71. Springer Cham (2021)
14. Golden, M.A., Picci, L.: Proposal for a new measure of corruption, illustrated with data. Econ. Polit. **17**(1), 37–75 (2005)
15. Montanari, G.E., Doretti, M., Bartolucci, F.: A multilevel latent Markov model for the evaluation of nursing homes' performance. Biom. J. **60**(5), 962–978 (2018)
16. OECD: Analytics for Integrity. Data-Driven Approaches for Enhancing Corruption and Fraud Risk Assessments. https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf (2019). Accessed 3 Jan 2022
17. Office Européen de Lutte Anti-Fraude (OLAF): Identifying and Reducing Corruption in Public Procurement in the EU. PwEU Service—Ecoyrs—Utrecht University (2013)

18. Pohle, J., Langrock, R., van Beest, F.M., Schmidt, N.M.: Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. J Agric. Biol. Environ. Stat. **22**(3), 270–293 (2017)
19. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
20. Syakur, M.A., Khotimah, B.K., Rochman, E.M.S., Satoto, B.D.: Integration K-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conf. Ser. Mater. Sci. Eng. **335**, 012017 (2018)
21. Thorndike, R.L.: Who belongs in the family. Psychometrika 267–276 (1953)
22. Transparency International: Transparency International. CPI Index. https://images.transparencycdn.org/images/CPI2020_Report_EN_0802-WEB-1_2021-02-08-103053.pdf (2020). Accessed 3 Jan 2022
23. Troìa, M.: Data analysis e costruzione di indicatori di rischio di corruzione per la Banca Dati Nazionale dei Contratti Pubblici. Autoritá Nazionale Anticorruzione ANAC, working paper no. 5 (2020)
24. Wiggins, L.M.: Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes. Elsevier, Amsterdam (1973)

A Multiplex Network Approach for Analyzing University Students' Mobility Flows

Ilaria Primerano, Francesco Santelli, and Cristian Usala

Abstract This paper proposes a multiplex network approach to analyze the Italian students' mobility choices from bachelor's to master's degrees. We rely upon administrative data on Italian students' careers by focusing on those who decide to enroll in a different university for their master's studies once they graduate in a bachelor's program. These flows are explored by defining a multiplex network approach where the ISCED-F fields of education and training are the layers, the Italian universities are the nodes, and the weighted and directed links measure the number of students moving between nodes. Network centrality measures and layers similarity indexes are computed to highlight the presence of core universities and verify if the network structures are similar across the layers. The results indicate that each field of study is characterized by its network structure, with the most attractive universities usually located in the Center-North of the country. The community detection algorithm highlights that graduates' mobility between universities is encouraged by the geographical proximity, with different intensities depending on the field of study.

Keywords Community detection · Layer similarity · Network centrality measures · Students' mobility

I. Primerano (✉)
Department of Political and Social Studies, University of Salerno, Fisciano, Italy
e-mail: iprimerano@unisa.it

F. Santelli
Department of Political Sciences, University of Naples Federico II, Naples, Italy
e-mail: francesco.santelli@unina.it

C. Usala
Department of Political and Social Sciences, University of Cagliari, Cagliari, Italy
e-mail: cristian.usala@unica.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_6

1 Introduction

Understanding the determinants and the main patterns of students' migration flows has become increasingly important to define the policies related to the attractiveness of the university systems in different areas [1]. Indeed, students' mobility is crucial since it affects several aspects related to regional competitiveness and growth capacity [2, 3] by also anticipating the future migration choices of highly educated individuals [4, 5]. These elements are particularly important if we consider the Italian case, characterized by substantial regional disparities in terms of educational outcomes, access to tertiary education, labor market conditions [6–8], and an unbalanced flow of students from southern to northern regions [9].

Several scholars have studied the determinants of students' mobility pathways by following different approaches and with different aims. In particular, some scholars have focused on the different factors affecting the mobility flows by investigating the role played by the hosting areas' characteristics [8], students' socio-economic and family background [10], peers' effects [11], and the quality of research [12]. These works have highlighted that universities' attractiveness positively depends on the quality of local amenities and job market opportunities, the quality of research and educational services.

Despite the growing interest in this phenomenon, most of the studies focused on the first level mobility (from high school to bachelor) [9, 13]. In contrast, only a few contributions have analyzed the second level mobility flows (from bachelor to master) by means of longitudinal analysis [14] and by focusing mainly on the choices of southern students [15]. Other studies have analyzed students' mobility flows by means of Social Network Analysis (SNA) [16] considering both international [17] and national levels [18], as well as aggregate data referred to macro-areas such as provinces and regions, and the specific field of study chosen by students [19]. A network perspective is considered to explore the structural patterns of students' mobility flows among Italian geographical areas focusing on the well-known South-North route [20]. Moreover, network measures, such as hub and authority scores, have also been used to analyze the direction and the intensity of student flows in a country and to label territories as good importers and/or good exporters [21]. These techniques have also been applied to reveal the presence of chain migrations phenomenons that link destination and origin geographical areas [22], and to analyze the mobility choices of Italian graduates that are choosing to attend a master's program [21].

Moving from this framework, an important element usually neglected is the presence of students' mobility flows between universities located within the same territory (e.g. city, province or region). In fact, even if a relevant part of the overall flows between universities in Italy is observed among institutions located in the same territory, most of the literature has focused only on the flows between macro areas.

In this paper, we aim to explore the role of Italian universities in the second level mobility network through a multiplex network approach [23]. In particular, students are considered in mobility when, after obtaining a bachelor's degree, they decide to

change their university to attend a master's degree program regardless of the distance between origin and destination universities.[1] Moreover, to highlight the differences existing among the fields of study, we define a multiplex network data structure by considering these latter as layers[2] [24], the Italian universities as the nodes, and the flows of students between universities as the weighted and directed links. Based on this network structure, we assess the similarities across fields by computing a set of layer similarities indexes [25] in order to highlight if there is a common pattern or if such fields are rather heterogeneous in terms of observed network structures.

The paper is structured as follows. Section 2 provides the definition and analysis of the multiplex mobility network. Section 3 describes the data used in the analysis and the normalization procedure adopted. Section 4 presents the main results, and concluding remarks are given in Sect. 5.

2 Multiplex Network Definition and Analysis

A multiplex network is a special case of multilayer networks [23, 26] where each layer holds a common set of nodes connected through two different kinds of relationships: intra-layer connections (i.e. edges linking nodes within the same layer) and inter-layer connections (i.e. edges crossing layers linking nodes in different layers).

Formally, let $\mathcal{M}$ be a multiplex network defined by a set of K different graphs $G(V, E_k)$, with $k = 1, \ldots, K$; where V is the set of common nodes and E_k is the set of both intra-layer edges (E_{kk}) and inter-layers edges (E_{kh}). For each layer k the corresponding adjacency matrix is $A_k = (a_{ijk})$, with $a_{ijk} = 1$ if $(v_i, v_j) \in E_k$, and $a_{ijk} = 0$ otherwise [23].

We define the layers of the Italian second level mobility network by grouping the degree programs in fields of study according to the ISCED-F classification. Since this latter is based on the similarities between programs in terms of disciplinary contents, it allows gathering in each layer the great part of the exchange flows of students between universities.

Thus, each observed layer holds a field-specific network where the set of nodes is represented by the Italian universities and the edges by the flows of students between them. In the second level mobility framework, intra-layer edges are given by the

[1] The definition of mobility adopted in this paper does not depend on the geographical distance between universities or territorial boundaries. For this reason, the Italian online universities are also included in the analysis.

[2] The educational fields are derived according to the 'International Standard Classification of Education: Fields of education and training' (ISCED-F) that was developed by the United Nations Educational, Scientific and Cultural Organization (UNESCO). In this contribution, the following ten broad fields are considered to define the layers of the multiplex network: Agriculture, forestry, fisheries and veterinary ('Agriculture'); Arts and humanities ('Arts'); Education; Engineering; manufacturing and construction ('Engineering'); Health and welfare ('Health'); Information and Communication Technologies ('ICTs'); Natural sciences, mathematics and statistics ('Sciences'); Services, Social sciences, journalism and information ('Social Sciences').

flows of students that change university after graduation but remain in the same field of study; while, inter-layer edges, consist of the flows of graduated students that change both university and field of study.

The resulting network is a weighted multiplex network [27], with the number of students involved in the flows defining the weights of both intra- and inter-layers edges. The edges of these networks are heavily influenced by universities' overall size which, in turn, depends on the number of students enrolled in the considered layers. Indeed, universities that provide programs in the most populated fields are more likely to have a higher number of students in mobility than those that provide programs in less chosen ones.

By construction, these networks are very rich in terms of edges (i.e. the networks are very dense) and thus tend to form a complete graph. Most connections likely consist of just a few students, while big universities will import and export a greater number of students because they can enroll more freshmen. For this reason, comparing layers holding universities with different sizes may not be appropriate since the most chosen universities are likely to rank both as top importers and top exporters in many fields. In order to get a simplified structure for each layer, a two-step procedure is applied. The first step normalizes the edges by defining a link-specific weight that depends on the size of the universities. In the second step, a cutting threshold value is set to identify the most relevant paths in the networks.[3]

Once all the layers have been processed, in depth analysis can be carried out to describe this structure through network comparisons in terms of similarity measures. Comparing networks is an important way to analyze multiplex network data structures, where the definition of similarity measures between layers is a key factor in appreciating their main characteristics. In its simplest form, this comparison consists of flattening all the layers into a single-layer network, where each actor is represented through a unique node linked by weighted edges whose weight depends on the number of layers on which the actors are connected. By contrast, the layer-by-layer methods process each layer separately and then compare the results of the SNA measures. Otherwise, ad hoc measures for multilayer networks have also been proposed by considering the difference between intra-layer and inter-layer edges and making a numerical distinction between layers or by analyzing edges involving different layers together but not mixing up with each other.

In particular, when dealing with layer similarity measures, several approaches have been proposed aiming to highlight the existence of the same structures across the layers and to identify their different characteristics [25]. A different approach has been developed for a visual exploration of the multiplex networks based on the use of factorial methods [28]. In this case, the authors analyzed the adjacency matrices derived by a multiplex network by using the DISTATIS technique [29]. This procedure allows showing the common structure of all layers (intra-layer perspective) and the variation of nodes across layers (inter-layer perspective).

[3] See Sect. 3 for details on the normalization procedure.

In order to quantify the shared characteristics of these structures across the layers, different similarity measures have been considered in terms of both actor-centered as well as layer-centered perspectives.

In this paper, a layer-by-layer perspective of analysis is applied [25] by focusing only on intra-layer edges to: (i) highlight the core universities for each field of study using network centrality measures [30]; and (ii) compare the results obtained across the layers by applying layer similarity measures.

Network centrality indexes have been computed to identify the universities that act as good importers or good exporters. Specifically, as for universities' attractiveness, we have considered the in-degree index, defined as the number of incoming edges. Instead, as a proxy of the export attitude of universities, we have computed the out-degree index, defined as the number of outgoing edges.

Furthermore, to compare the results across the layers, we take advantage of two indexes: the Pearson Degree Similarity coefficient [31] and the Jaccard Layer Correlation Coefficient [25]. The Pearson Degree Similarity Coefficient quantifies the similarity among nodes' degrees across layers and allows to assess universities' centrality across different disciplinary fields. The Jaccard Layer Correlation Coefficient measures the overlapping between pairs of layers. It varies between 0 and 1, with 0 indicating no overlapping and 1 indicating perfect overlapping. This coefficient highlights the presence of edges among the same nodes on different layers, i.e. the presence of edges linking the same group of universities in different disciplinary fields. It is used to identify the presence of common structures between fields and, at the same time, to determine the turnover that takes place between the layers.

Finally, in each field of study, the Clauset-Newman-Moore (CNM) community detection algorithm [32] has been adopted to identify the presence of groups of universities which are tightly connected in knit groups within communities and loosely connected with universities belonging to other communities.

3 Data Description and Normalization Procedure

Italian students' second level mobility network is constructed starting from the database MOBYSU.IT [33] that holds the administrative data regarding all Italian university students.[4] In particular, this database includes information on students' careers that allows us to identify the universities in which the students have obtained their bachelor's degrees and those chosen for their master's studies.

We extracted the data regarding the cohorts of Italian students who started their university careers in a bachelor's program between a.y. 2011/2012 and a.y. 2015/2016, and have enrolled in a master's degree program between a.y. 2014/2015

[4] Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Universitá e Ricerca', and INVALSI.

Table 1 Distribution of students in mobility according to ISCED-F fields

Code	ISCED-F 2013	Master's students	Students in mobility	Same field
		N.	N. (%)	(%)
L1	Agriculture, forestry, fisheries and veterinary	10,959	2,631 (24.0%)	65.8
L2	Arts and humanities	63,756	19,440 (30.4%)	72.4
L3	Business, administration and law	59,949	15,849 (26.4%)	78.1
L4	Education	12,728	3,245 (25.4%)	83.4
L5	Engineering, manufacturing and construction	82,753	12,319 (14.8%)	94.3
L6	Health and welfare	11,258	4,309 (38.2%)	61.1
L7	Information and Communication Technologies (ICTs)	4,339	816 (18.8%)	80.6
L8	Natural sciences, mathematics and statistics	41,195	10,945 (26.5%)	93.1
L9	Services	12,484	3,979 (31.8%)	37.0
L10	Social sciences, journalism and information	68,304	23,713 (34.7%)	75.4

and a.y. 2018/2019. Moreover, since students can enroll in different universities during their careers, in the network definition we consider the origin university the one in which she/he has obtained her/his last bachelor's degree. Namely, if a student has gained two or more bachelor's degrees in her/his career, we consider as the origin node only the last university in which the student has been observed. Then, since students may also enroll in several master's programs, we consider as destination node only the first university in which the student has started her/his master's programs.

Therefore, starting from the population of 1,171,006 Italian bachelor's students, the dataset holds information about 621,075 (53%) students that have graduated in the time frame considered. Secondly, we keep the information regarding only those students that have enrolled in a master's program. The analysis is based on 367,725 students, belonging to 92 universities (of which 11 are online universities).

Table 1 presents the descriptive statistics of the distribution of students among fields of study. In particular, for each ISCED-F field, the table shows the number of graduated students enrolled in a master's program in Italy (Master's students); the number and the percentage of students that have changed university for attending a master's program (Students in mobility); and the percentage of students in mobility that have chosen to stay in the same field of study (same field).

Table 1 shows that the two most chosen groups are 'Engineering' and 'Social sciences' while the least chosen is 'ICTs'. Regarding the second level mobility, we observe that at least 14.8% of students in each field have changed university after

graduation, with a maximum of 38.2% in 'Health'. Considering the last two columns, it is noticeable that the tendency to change the field of study varies depending on the field considered, with the 94.3% of 'Engineering' graduates that have not changed their field after graduation, whereas the majority of students in the 'Services' field (63.0%) have decided to enroll in a different field.

In order to analyze the second level mobility networks we apply a two-step procedure. The first step consists in normalizing the weights given by the number of students involved in the flows. Following Slater [34], we apply a Multidimensional Iterative Proportional Fitting Procedure (MIPFP) [35]. This procedure sets as seed the original adjacency matrix, where the rows' and columns' marginals are, respectively, the total number of incoming and outgoing students for a given university, and then performs the reshaping. The set of known desired target marginal counts M is a non-empty subset (2–dimensional) of:

$$\tau = \left\{(a_{\bullet j}), (a_{i\bullet})\forall i, j\right\} \tag{1}$$

where $\bullet$ is the summation over the corresponding university. The procedure iteratively updates the values of the cells depending on the targets. The first iterations for both marginals are:

$$a_{ij}^1 = a_{ij}^0 \cdot \frac{a_{\bullet j}}{a_{\bullet j}^0} \quad \forall i, j \quad ; \quad a_{ij}^2 = a_{ij}^1 \cdot \frac{a_{i\bullet}}{a_{i\bullet}^1} \quad \forall i, j \tag{2}$$

The adjustments at a generic iteration l stop when the stopping criterion given by the *tol* parameter (a small constant) is reached:

$$\max \left| a_{ij}^{l-1} - a_{ij}^{l} \right| \leq tol \quad \forall i, j \tag{3}$$

This procedure defines a value for each edge ranging from 0 to 1 accounting for nodes' attractiveness (columns marginal) and nodes' export attitude (rows marginal). Thus, at the end of the normalization, each weight is a value that takes into account both the overall number of incoming and outgoing edges. Namely, edges' weights inversely depend on the number of students: higher weights are associated to universities with a small number of flows; lower weights are attached to universities characterized by a relatively large number of incoming and outgoing students.

Since the normalized network is very dense, a cut-off threshold is set at the median value of the non-zero normalized entries in the second step. This value is used to dichotomize the obtained weights in order to let the most relevant flows emerge.

4 Main Results

The multiplex mobility network under analysis consists of 10 layers and 5.689 total intra-links. Table 2 shows the descriptive statistics on the main structural characteristics of the networks. Some network measures have been computed at the network level for each layer and the flattened network. In particular, the table shows the number of actors (n), the number of edges (m), the number of strong components (nc), the density (den), and the clustering coefficient (cc). For all the networks, given the normalization procedure, a low value of the density is observed, with maximum values of 0.13 for both 'Health' and 'Social Sciences' and a minimum value of 0.06 for 'Agriculture'. At the same time, the global clustering coefficient is almost homogeneous across the layers, ranging from 0.31 to 0.46.

Moreover, to highlight the role played by each university in all fields of study, we have computed the centrality measures for every single layer and the flattened network. In particular, in-degree and out-degree centrality measures have been computed to identify the most attractive universities (i.e. high in-degree value) and those losing students (i.e. high out-degree value).

Concerning the flattened network, the top five universities in terms of attractiveness are Pisa, Bologna, Florence, Milan, and Turin. In contrast, the top five exporters are Parma, Catania, Pisa, Florence and Modena. Although some of these are so-called big universities, for which it is customary to expect a larger number of enrolled students, it is worth noting that the two-step normalization procedure takes into account

Table 2 Descriptive measures on student mobility multiplex network by fields of study

Layers		n	m	nc	den	cc
	Flattened network	92	5,689	1	0.68	0.66
L1	Agriculture, forestry, fisheries and veterinary	52	168	2	0.06	0.35
L2	Arts and humanities	88	868	1	0.11	0.31
L3	Business, administration and law	87	827	1	0.11	0.32
L4	Education	57	251	1	0.08	0.46
L5	Engineering, manufacturing and construction	81	498	2	0.08	0.35
L6	Health and welfare	71	663	1	0.13	0.40
L7	Information and Communication Technologies (ICTs)	79	562	1	0.09	0.32
L8	Natural sciences, mathematics and statistics	62	387	1	0.10	0.41
L9	Services	75	490	2	0.09	0.46
L10	Social sciences, journalism and information	80	835	1	0.13	0.40

the overall size of universities. In fact, the flattened network results show the presence of smaller universities as well.

As regards single-layer networks, the in-degree centrality results show that the top five universities are very different among the 10 layers considered, indicating that each field is characterized by its peculiar structure. Specifically, few universities are present in more than one layer. For example, Pisa ranks in the top positions in five different layers: 'Arts', 'Education', 'Engineering', 'ICTs', and 'Social sciences'. It is also remarkable that in the top positions for some layers, we find several online universities (UNITEL, Niccoló Cusano, Pegaso, and UniNettuno) and universities located in metropolitan areas (e.g. Turin, Milan, and Rome).

The out-degree results also emphasize the heterogeneity among the layers, but involving different groups of universities. Indeed, the out-degree values reflect the well-known pattern of Italian student migration, with many outgoing flows originating in Southern Italy universities that are directed towards the universities of central and northern Italy.

Considering, for instance, the two most chosen fields by Italian mover students, 'Social Sciences' and 'Engineering', the top five attractive universities in 'Social Sciences' field are IULM of Milan, Carlo Bo of Urbino, Perugia, Pisa and Siena, while the top five exporters are the universities of Bologna, Brescia, Modena, Parma and Pavia. Instead, the mobility flows in the 'Engineering' field show as the top five attractive universities La Sapienza of Rome, Bologna, Milan, Pisa, and Turin, while those losing students are Bicocca of Milan, Padua, Salerno, Trento and Udine.

Overall, the single-layer analysis shows that it is uncommon for a university to be in a leading position in all the fields. Indeed, the high position of a university in the ranking of the most attractive ones could be related to the presence of a 'department of excellence' that could act as a pole of attraction for students, making the university to which they belong a central node in that specific layer. What is instead very interesting is the centrality of the online universities in different layers in which they act primarily as importers rather than exporters in the second level mobility framework.

Moving to the multiplex mobility network analysis, the two similarity measures described in Sect. 3 have been computed to compare layers' properties.

For what concerns the Pearson Degree Similarity coefficient, reported in Table 3, the higher values are observed for the 'ICTs' and other fields: 'Agriculture' (0.74), 'Business' (0.65), and 'Social Sciences' (0.60). Other high correlations involve 'Services' with 'Arts' (0.63) and 'Social sciences' (0.63) fields. With respect to the Jaccard coefficient, it is noticeable that all the values reported in Table 4 are very close to 0. These results perfectly align with what emerged in the single-layer analysis, where the central actors differ among the observed layers. In other words, the mobility flows considered, rather than defining a common set of core universities for all the ISCED-F fields, highlight the presence of different typical structures in each specific field.

In the following, the results obtained for a specific layer are presented to show the application scope of our analysis procedure. Specifically, we consider the normalized network of the layer 'Education' whose representation is given in Fig. 1.

Table 3 Correlation matrix of the Pearson Degree Similarity coefficient by fields of study

		L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
L1	Agriculture	1.00	–	–	–	–	–	–	–	–	–
L2	Arts	0.56	1.00	–	–	–	–	–	–	–	–
L3	Business	0.41	0.52	1.00	–	–	–	–	–	–	–
L4	Education	0.57	0.28	0.35	1.00	–	–	–	–	–	–
L5	Engineering	0.54	0.41	0.41	0.47	1.00	–	–	–	–	–
L6	Health	0.45	0.34	0.51	0.39	0.41	1.00	–	–	–	–
L7	ICTs	0.74	0.56	0.65	0.50	0.49	0.57	1.00	–	–	–
L8	Sciences	0.51	0.40	0.47	0.33	0.50	0.33	0.49	1.00	–	–
L9	Services	0.38	0.63	0.58	0.23	0.38	0.43	0.49	0.46	1.00	–
L10	Social sciences	0.55	0.51	0.55	0.32	0.43	0.48	0.60	0.46	0.63	1.00

Table 4 Correlation matrix of the Jaccard edge similarity coefficient by field of study

		L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
L1	Agriculture	1.00	–	–	–	–	–	–	–	–	–
L2	Arts	0.15	1.00	–	–	–	–	–	–	–	–
L3	Business	0.09	0.11	1.00	–	–	–	–	–	–	–
L4	Education	0.10	0.06	0.07	1.00	–	–	–	–	–	–
L5	Engineering	0.08	0.05	0.06	0.08	1.00	–	–	–	–	–
L6	Health	0.11	0.10	0.12	0.06	0.06	1.00	–	–	–	–
L7	ICTs	0.16	0.15	0.12	0.08	0.08	0.12	1.00	–	–	–
L8	Sciences	0.12	0.11	0.09	0.05	0.07	0.09	0.12	1.00	–	–
L9	Services	0.13	0.18	0.10	0.06	0.04	0.10	0.13	0.12	1.00	–
L10	Social sciences	0.14	0.17	0.10	0.06	0.04	0.11	0.13	0.11	0.17	1.00

This graph shows the edges connecting the 57 Italian universities among which the most important exchange flows occur. The size of each node indicates the level of attractiveness of the university in this field, measured through the in-degree centrality index. Moreover, nodes are colored according to the cluster (i.e. communities) they lie in. In this layer six communities have been identified by means of The Clauset-Newman-Moore (CNM) community detection algorithm.

By exploring the patterns outlined by the flows of students who change university for their master's studies while remaining in the Education field, the geographical aspect of mobility clearly emerges. This element is mainly due to the fact that students also move into the same geographical boundaries. In fact, looking at the six communities, this aspect is evident. The group on the top-left of Fig. 1 (dark green) consists of the universities that are located in the central and southern regions of the eastern part of Italy, such as Apulia, Molise, Abruzzo and Marche. Some of these universities, i.e. Foggia and Bari, have many incoming and outgoing flows. Instead,

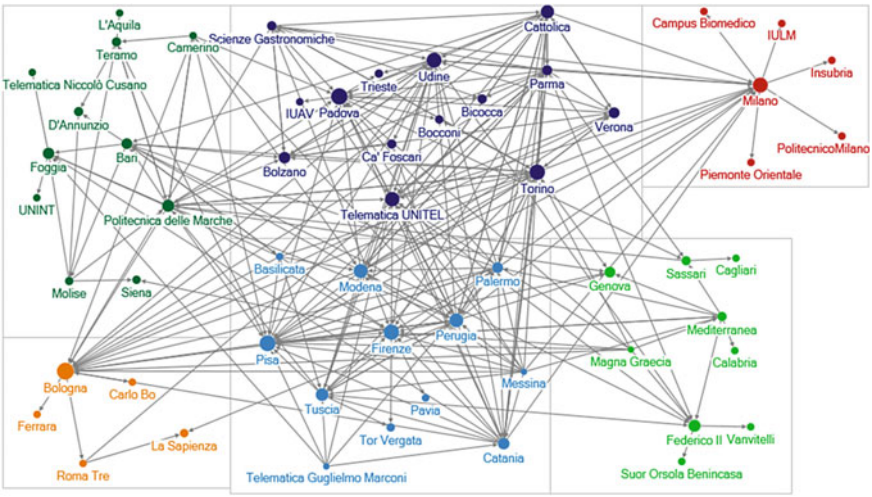


Fig. 1 Single layer visualizations of the student mobility network for the education field of study. Nodes' colors depend on the community; nodes' size depends on the in-degree centrality index

the Politecnica of Marche has a strong attitude toward attracting students, even from universities located outside the community.

The first community (orange), on the bottom-left of the Fig. 1, has Bologna university as the attractive pole of the community that also includes other universities from Emilia-Romagna, Lazio, and the Urbino University.

Indeed, on the top-right, the second community (red) groups most of the universities located in the Lombardy region, whose edges form a star-shaped configuration around the university of Milan. Thus, in this community, the centrality role of the university of Milan clearly emerges. At the same time, the university of Milan acts as a local hub by attracting students from all the other regions and yielding a quota of its graduates to other universities in the same geographic area.

The third community (green), on the bottom-right, is mainly composed of the universities located in Campania, Sardinia, and Calabria. This is a dense community, with many flows within it, in which emerges the central role of the Federico II University of Naples characterized by many incoming flows.

Finally, the two big communities in the center of Fig. 1 are related to North universities (top-center, dark blue color) and Center and South universities (top-center, light blue color). These communities show many incoming and outgoing flows without a specific single region involved but showing high values for both in-degree and out-degree indexes.

As the last remark, Fig. 2 shows, for each field of study, the skeleton graph of the communities, where each community is a vertex. This visualization highlights the relationships among different communities (i.e. flows of students among communities) and the isolated ones characterized only by internal flows of students (i.e. flows of students within the same community). Finally, the size of the vertexes is

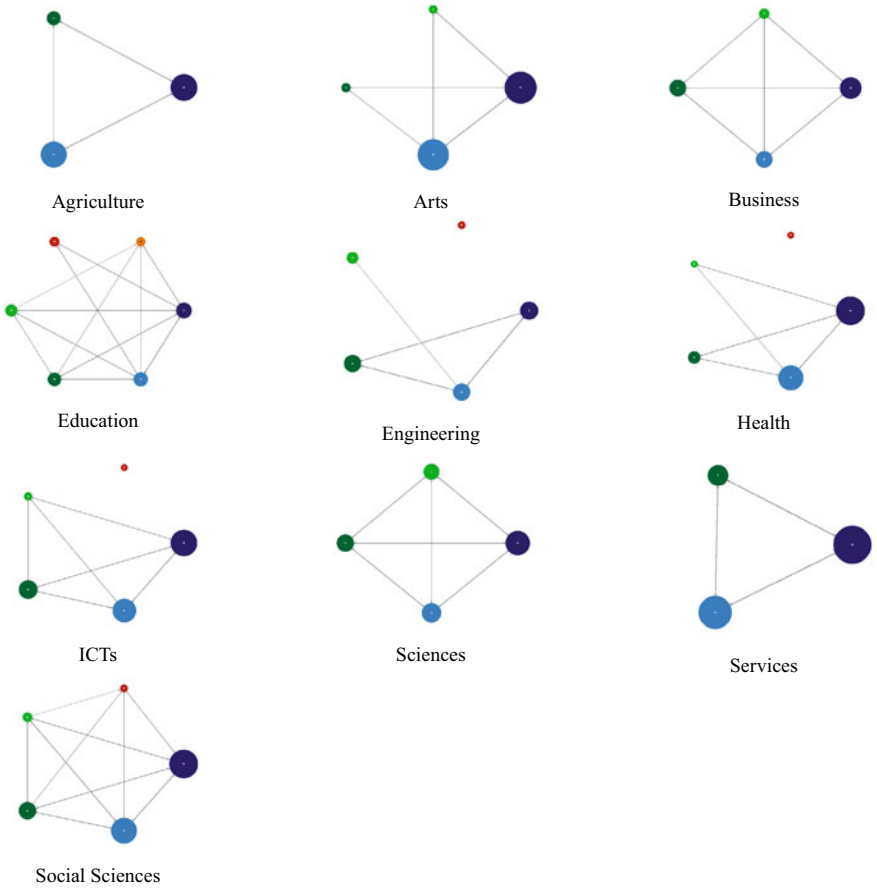


Fig. 2 Skeleton graph representation of the communities detected for each field of study

proportional to the number of internal flows. The skeleton representation in Fig. 2 of the six communities belonging to the field 'Education' allows us to notice that the first and the second communities are reached by all the other ones, thus meaning that these communities include the most attractive universities for the considered layer. Indeed, the second community holds the university of Milan, while the first the one of Bologna, which are the two most popular universities in this layer.

5 Concluding Remarks

In this contribution, we have introduced the study of second level student mobility into the framework of multiplex network analysis. Our results show clear hints on the students' flows at the moment of enrollment in a master's degree program. Such kind of mobility has not been explored deeply in the Italian context. To the best of our knowledge, this is the first study on the Italian second level mobility that accounts, at the same time, for several aspects: including online institutions, considering the differences among fields of study, and neglecting the geographical setting.

Furthermore, in line with previous studies [21, 36], the evidence obtained through the community detection analysis shows that one of the most important factors that encourage second level mobility is the geographic proximity. Moreover, most attractive universities are located in the northern and central regions. However, such geographical influences on students' flows have a different intensity depending on the field of study considered. In fact, the layer comparison analysis has clearly shown that the layers differ in terms of their network structures and their flow dynamics. This element may be due also to the fact that the supply of degree programs in Italy is heterogeneous and some universities provide only a few programs in specific fields.

Several aspects could be considered when dealing with students' mobility data organized into multiplex structures, such as the analysis of the inter-layers connections to understand the determinants of students' decisions to change their field after graduation. Moreover, future lines of research include the study of the relationship between the overall universities' attractiveness scores, their supply of educational services and hosting areas' characteristics, and the assessment of the sensitivity of the results to different normalization procedures and cutting thresholds.

Acknowledgements This contribution has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide", n. 2017HBTK5P—CUP B78D19000180001.

References

1. Attanasio, M.: Students' university mobility patterns in Europe: an introduction. Genus **78**(17) (2022)
2. Krugman, P.: Increasing returns and economic geography. J. Polit. Econ. **99**(3), 483–499 (1991)
3. Valero, A., Van Reenen, J.: The economic impact of universities: evidence from across the globe. Econ. Educ. Rev. **68**, 53–67, 68 (2019)
4. Dotti, N.F., Fratesi, U., Lenzi, C., Percoco, M.: Local labour markets and the interregional mobility of Italian university students. Spat. Econ. Anal. **8**(4), 443–468 (2013)
5. Oggenfuss, C., Wolter, S.C.: Are they coming back? The mobility of university graduates in Switzerland. Rev. Reg. Stud. **39**, 189–208 (2019)
6. Türk, U.: Socio-economic determinants of student mobility and inequality of access to higher education in Italy. Netw. Spat. Econ. **19**(1), 125–148 (2019)

7. Fratesi, U., Percoco, M.: Selective migration, regional growth and convergence: evidence from Italy. Reg. Stud. **48**(10), 1650–1668 (2014)
8. Giambona, F., Porcu, M., Sulis, I.: Students mobility: assessing the determinants of attractiveness across competing territorial areas. Soc. Indic. Res. **133**(3), 1105–1132 (2017)
9. Attanasio, M., Enea, M., Priulla, A.: Quali atenei scelgono i diplomati del Mezzogiorno d'Italia? Neodemos, ISSN: 2421-3209 (2019)
10. Impicciatore, R., Tosi, F.: Research in social stratification and mobility student mobility in Italy: the increasing role of family background during the expansion of higher education supply. Res. Soc. Stratif. Mobil. **62**, 100409 (2019)
11. Porcu, M., Sulis, I., Usala, C.: Estimating the peers effect on students' university choices. In: Lombardo, R., Camminatello, I., Simonacci, V. (eds.) Book of Short Papers IES 2022: Innovation & Society, pp. 134–139 (2022)
12. Bratti, M., Verzillo, S.: The 'gravity' of quality: research quality and the attractiveness of universities in Italy. Reg. Stud. **53**(10), 1385–1396 (2019)
13. D'Agostino, A., Ghellini, G., Longobardi, S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electron. J. Appl. Stat. Anal. **12**(4), 826–845 (2019)
14. Enea, M.: From south to north? Mobility of southern Italian students at the transition from the first to the second level university degree. In: Perna, C., Pratesi, M., Ruiz-Gazen, A. (eds.) Studies in Theoretical and Applied Statistics, pp. 239–249 (2018)
15. Attanasio, M., Enea, M., Albano, A.: Dalla triennale alla magistrale: continua la 'fuga dei cervelli' dal Mezzogiorno d'Italia. Neodemos, ISSN: 2421-3209 (2019)
16. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
17. Restaino, M., Vitale, M., Primerano, I.: Analysing international student mobility flows in higher education: a comparative study on European countries. Soc. Indic. Res. **149**(3), 947–965 (2020)
18. Santelli, F., Scolorato, C., Ragozini, G.: On the determinants of students mobility in an interregional perspective: a focus on Campania region. J. Appl. Stat. **31**(1), 119–142 (2019)
19. Columbu, S., Primerano, I.: A multilevel analysis of university attractiveness in the network flows from bachelor to master's degree. In: Pollice, A., Salvati, N., Schirippa Spagnolo, F. (eds.) Book of Short Papers SIS 2020, pp. 480–485 (2020)
20. Vitale, M., Giordano, G., Ragozini, G.: University student mobility flows and related network data structure. In: Pollice, A., Salvati, N., Schirippa Spagnolo, F. (eds.) Book of Short Papers SIS 2020, pp. 515–520 (2020)
21. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Analysing the determinants of Italian university student mobility pathways. Genus **78**(6) (2022)
22. Genova, V.G., Tumminello, M., Enea, M., Aiello, F., Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electron. J. Appl. Stat. Anal. **12**(4), 774–800 (2019)
23. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Complex. Netw. **2**, 203–271 (2014)
24. UNESCO Institute for Statistics: ISCED Fields of Education and International Standard Classification of Education 2011. Montrèal (2014)
25. Bródka, P., Chmiel, A., Magnani, M., Ragozini, G.: Quantifying layer similarity in multiplex networks: a systematic study. R. Soc. Open Sci. **5**, 171747 (2018)
26. Dickison, M.E., Magnani, M., Rossi, L.: Multilayer Social Networks. Cambridge University Press (2016)
27. Menichetti, G., Remondini, D., Panzarasa, P., Mondragón, R.J., Bianconi, G.: Weighted multiplex networks. PloS One **9**(6), e97857 (2014)
28. Giordano, G., Ragozini, G., Vitale, M.P.: Analyzing multiplex networks using factorial methods. Soc. Netw. **59**, 154–170 (2019)
29. Abdi, H., Dunlop, J.P., Williams, L.J.: How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage **45**(1), 89–95 (2009)

30. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978)
31. Berlingerio, M., Coscia, M., Giannotti, F.: Finding and characterizing communities in multidimensional networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE, pp. 490–494 (2011)
32. Clauset, A., Newman, M.E., Moore, C: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
33. Database MOBYSU.IT [Mobilitá degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenza, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA
34. Slater, P.B.: Multiscale network reduction methodologies: bistochastic and disparity filtering of human migration flows between 3,000+ us counties (2009). arXiv:0907.2393
35. Barthélemy, J., Suesse, T.: mipfp: an R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. J. Stat. Softw. **86**(2) (2018)
36. Santelli, F., Ragozini, G., Vitale, M.P.: Assessing the effects of local contexts on the mobility choices of university students in Campania region in Italy. Genus **78**(5), 1–25 (2022)

A Statistical Model for Predicting Child Language Acquisition: Unfolding Qualitative Grammatical Development by Using Logistic Regression Model

Andrea Briglia, Massimo Mucciardi, and Giovanni Pirrotta

Abstract Language acquisition is a scientific puzzle still awaiting a theoretical solution. Children seem to acquire their native language in a spontaneous and effortless way and they probably do so by keeping track of the frequency with which language items such as phonemes or parts of speech occur. Advances in data storage, processing and visualization have triggered a growing and fertile interest in analysing language by relying on statistics and quantitative methods. In this paper we propose a multiple logistic regression model to evaluate how different components of language contribute to its acquisition over time. The empirical basis consists of a corpus, which can be considered as a series of statistically representative samples taken at regular time intervals. The aim is to show how quantitative methods can contribute to explaining the creation and development of grammatical categories in first language acquisition.

Keywords Natural language processing · Multiple logistic regression · Phonetic variation · Frequency effects on learning

1 Introduction

This paper is to be considered as a continuation of a previous research project [3, 16, 17] in which the phonetic development of children was explored. In the current paper we have extended the level of analysis from a merely phonetic one to give a more global view on how phonemes turn into words. The elementary units are Part Of

A. Briglia
STIH Lab LC Sorbonne Université, Paris, France
e-mail: andrea.briglia@sorbonne-universite.fr

M. Mucciardi (✉)
Department of Cognitive Science, University of Messina, Messina, Italy
e-mail: massimo.mucciardi@unime.it

G. Pirrotta
University of Messina, Messina, Italy
e-mail: giovanni.pirrotta@unime.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_7

Speech tags (from now POS tags). If phonemes could be metaphorically considered as the atoms of language, in a similar fashion, words could be viewed as playing the role of the molecules: though the latter are far bigger than the former, they combine in different ways to form more complex meaningful entities in an analogous manner. Children always need to infer rules and regularities of their native language from a limited amount of input. Their task requires them to: «[..] discover the underlying structure of an immense system that contains tens of thousands of pieces, all generated by combining a small set of elements in various ways. These pieces, in turn, can be combined in an infinite number of ways, although only a subset of those combinations is actually correct. However, the subset that is correct is itself infinite. Somehow you must rapidly figure out the structure of this system so that you can use it appropriately early in your childhood» [19].

Learning a language means learning how to creatively combine a set of units that must respect a conventional order (i.e., phonotactic constraints and grammar): to reach this cognitive ability, children do not acquire their native language by simply repeating the input received; doing it this way would require much more time than observed. Children actively optimize the input received by trying to check whether it fits with adult language: for example, English speaking children often pronounce an irregular verb conjugation by adding the regular "-ed" suffix on it. As "-ed" is the most frequent suffix, whenever a child does not know how to conjugate a verb that he has never heard, the best option is to not take any risk and treat it as if it was part of the most common category.

This is due to the fact that "high frequency forms are early acquired and prevent errors in contexts where they are the target, but also cause errors in contexts in which a competing lower-frequency form is the target" [1].

As a further example, Italian speaking children do not need to hear all verbs listed in the "-are", "-ere", "-ire" forms to be sure how to conjugate the corresponding suffixes: once they know which rule applies to the different singular and plural forms of each person, they become able to apply this rule even to verbs that they have never heard (meaning the vast majority) [1].

This flexibility is the core of human learning: to infer as much information as possible in the most reliable way from a sample of reality. It could be said that human minds have been shaped by evolutionary pressure to be inference engines, able to extract the best sample of the world through the better optimized sampling method [8].

Modeling first language acquisition is a challenging scientific puzzle at the core of human cognition. Recent advances in Natural Language Processing focus on modelling language through Neural Network. Sagae [20] proposes a state-of-the-art example that has been trained on the same type of linguistic data we used for our study (i.e. part of the CHILDES project [14]). The American author says that "with the development of computational models for syntactic analysis of child language utterances, automatic accurate computation of syntax-based metrics of language development became possible" [20]. More traditional and well-established approaches are word vectorization (known as "word2vec") as well as word probability distribution models following a power-law probability distribution (also known

as a Pareto-like or Zipfian-like distribution) [7]. A fairly recent and promising field in first language acquisition is the application of Bayes' rule of conditionalization on learning processes regarding perception, language[1] or other cognitive abilities [5].

There are several stages in first language acquisition, ranging from cooing and babbling to the mastery of long and embedded sentences containing abstract references. This covers a time span starting from birth until 5–6-year-old. As grammar could be defined as the set of rules structuring the morphological and syntactical aspects of any given language, it could then be said that it starts to develop from the two-word stage [21], despite the fact that inflectional morphology (gender and plural forms, depending on language) is already at play in one word utterance.

Modelling first language acquisition is a challenging scientific puzzle impossible to tackle in a paper: what is at stake here is proposing a multiple logistic regression model that has shown good performance in predicting grammatical development. The paper is organized as follows: in Sect. 2 we present the data optimization strategy, before presenting in Sect. 3 the statistical model proposed and the main results obtained. Finally, Sect. 4 provides conclusions and suggestions for future research.

2 Sampling and Data Optimization Strategy

A key element in child language acquisition is sampling. The core question that needs to be addressed is "how much to sample and at what intervals and for how long and for how many children" [23]. In other words, what does a statistically representative sample need to have in this specific field? It is clear that reduced or low-density sampling can give good results if the target structure is highly frequent, but if the target is a rare phenomenon, the same sampling technique would become insufficient. In addition, time constraints, the lack of long-lasting funding opportunities and external reasons often impede researchers from obtaining an ideal corpus.

There are two different ways of collecting data to study language acquisition: longitudinal and cross-sectional studies.

- Longitudinal studies capture the continuous language development of one child, and the premise is that this individual development might be generalized to the global language development of children who speak this particular language.
- Cross-sectional studies captures stages of language development in children of different ages, and the premise is that these different stages might represent a continuous temporal development [2, 25].

For this paper we will refer to the first type of sampling and for this reason we use the database CoLaJE [15] where temporal density of sampling is adequate.

CoLaJE is an open access French database part of the broader CHILDES project [14]: seven children have been recorded in a natural setting one hour every month, from their first year of life until approximately five years of age. Data is available in three different formats: International Phonetic Alphabet (IPA), orthographic norm and Code for the Human Analysis of Transcription (CHAT), each of which is aligned

Loc	Ts	Te	Transcription L: 990 - 988 - 997 T: (-) P: - 0:00:09
CHI	0:24:03	0:24:13	‹regarde celui-là là il a trois essuie-glace› !
pho			‹ʁəgaʁ sɥila la il a kwa sysygas›
mod			‹ʁəgaʁd sɥila la il a tʁwa ɛsɥiglas›
FAT	0:24:13	0:24:14	il a trois essuie-glace .
CHI	0:24:14	0:24:15	oui !
pho			wi
mod			wi
FAT	0:24:14	0:24:16	essuie-glace Adrien , comment tu dis ?
FAT	0:24:16	0:24:19	tu dis essuie-glace ?
CHI	0:24:19	0:24:21	‹celui-là il en a deux› !

Fig. 1 Extract from CoLaJE French database

to the correspondent video recording, allowing researchers to see the original source and to eventually reinterpret every utterance on their own. The main coding structure of the database consists in the fundamental division between "*pho*" (what the child says) and "*mod*" (what the child should have said according to the adult standard phonetic/phonological norm): we define every occurrence in which "*pho*" differs from "*mod*" as a variation (An example of the database is shown in Fig. 1).

The sampling scheme can influence the range of deductions and generalizations that we can draw from the data. For this reason, we checked if this corpus sampling scheme meets internationally recognized reliability criteria [23], which it does: this means that it is considered as statistically representative with respect to the frequency of the linguistic structures targeted.

We transformed all the sentences for the child named *Adrien*[1] from 2 to 5 years old (8214 sentences, 19093 words) to machine-readable strings of characters in order to make them computable by the Python STANZA library [18] software. STANZA library features a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech, morphological feature tagging, dependency parsing, and named entity recognition. We have chosen STANZA package because it has been proved to be a state-of-the-art automatic POS tagging system, as described in a more detailed way in an article [18] in which the neural pipeline is compared to competitors.

We tagged all the words of the sentences by assigning a part-of-speech (POS) tag to each. In the second step, we calculated the "Word Phonetic Variation" (WPV) for each word by setting a specific algorithm to compute this difference. At this step, we assume that a correct word is a word that has been correctly pronounced though we are aware that—grammatically speaking—a word needs also to be pronounced in the correct place to be considered fully correct (i.e. in the correct order). This is the case for the majority of sentences (especially shorter ones). This assumption is considered

[1] Among the seven children in the CoLaJE database we choose Adrien because, from a sampling point of view, the data is more detailed and complete.

acceptable because programming a set of grammar-sensitive and context-dependent algorithms is a hard challenge, especially for evolving linguistic structures such as those of children, besides not to mention the fact that every language has its own specific grammar. In other words, this analysis has been made independently of the cardinality of the original sentences to which the words belong. We then organize the data in a spreadsheet structure on which we have built the statistical model.

3 Statistical Model and Variables Selection

From a statistical point of view, we developed a multiple logistic regression model to examine which factors can predict a child's performance as part of our methodology.[2] As we know, logistic regression is a statistical procedure developed for binary data. It generally describes the probability of a 0 or 1 outcome with an S-shaped logistic function [9].

The binary variable was set as follows: WPV = 1 if there is a phonetic variation in the spoken word and WPV = 0 if there is no phonetic variation.

We choose 4 predictors to explain WPV: AGE, COMPLEX, IPC and CLASS.

- AGE is the main driver of development: the more a child is exposed to his environment, the more he will learn from it. It represents *Adrien*'s age from 1 to 5 years[3];
- COMPLEX relates to the difficulty a child has to master long and semantically rich sentences where more cues need to be spotted. It represents the type/token ratio, meaning the percentage of distinct words per sentence (i.e. a proxy of lexical richness). Moreover it represents the complexity of a sentence, not just by its length but as well as by its richness;
- IPC (Index of Phonetic Complexity). This index has been developed by American linguist Katy Jakielski [11] and has been adapted to French language [12]. It is composed by eight parameters which determine a final score that represents the articulatory difficulty of any given word of a specific language. The parameters seek to analyze phonetic complexity according to the "frame/content theory of evolution of speech" [13]. A point is attributed to a word if it contains one of the eight parameters. Here are some examples: a word ending with a consonant will receive a point, whereas a word ending with a vowel does not; a word containing a consonantal cluster such as "tr" will receive a point, a word having a simpler consonant/vowel structure will score zero and so on (consult the original article [12] to read the full list of parameters).[4] It is important to highlight that the difference between IPC and COMPLEX is in the level of the analysis: IPC represents the phonetic (or intra-word) level while COMPLEX represents the syntactic (or

[2] Classical linear regression was discarded because it gave poor results during the modelling stage.

[3] We transformed the variable from months to years for a better representation of the data.

[4] It is possible to calculate the IPC of any given French word at this link [10]: http://igm.univ-mlv.fr/~gambette/iPhocomp/.

Table 1 Main statistics for a spreadsheet structure

Variables	%	*Mean*	*SD*
WPV (Yes)	31.7	–	–
WPV (No)	68.3	–	–
AGE	–	3.53	0.63
COMPLEX	–	0.84	0.20
IPC	–	1.10	1.59
Class (Open)	53.8	–	–
Class (Closed)	45.2	–	–
Class (Other)	1.0	–	–
Sample size = 19,093 words from 8214 sentences			

supra-word) level. These are two models of a unique reality: of course, phonemes and sentences are part of the same entity. Here we represent them by using two levels of analysis to target in a more focused way the respective contribution of phonetics and syntax in first language acquisition.

– CLASS provides a way to evaluate whether a child can or cannot use a given grammatical element, giving an indirect measure of his grammatical development. This predictor works by specifying the class to which the word belongs: Open, Closed or Other. "Open" contains lexical words such as verbs, common and proper nouns, and adjectives. These classes contain large numbers of elements and are subject to change (a new entry can be added, another can be deleted). "Closed" contains functional words such as auxiliaries, pronouns and determiners. These classes contain few but highly occurring elements that are not subject to change (no new entries at all). "Other" contains everything that cannot be classified in the previous categories (punctuation, acronyms, etc.).

This framework of three classes has been taken from the Universal Dependencies project [24].[5]

The main statistics for the dependent variable and for the predictors are summarized in Table 1.

The following logit Eq. (1) will give the probability of WPV based on the four regressors:

$$P(Y|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 AGE+\beta_2 COMPLEX+\beta_3 IPC+\beta_4 CLASS)}} \quad (1)$$

Based on the equation analysis results,[6] we can see which variable among AGE, COMPLEX, IPC and CLASS variables are statistically significant (Table 2). The

[5] Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages.

[6] Before modelling AGE as linear, we tried to create three successive yearly time slots to see how the two other regressors behave if taken apart, but the resulting classified number of cases had a lower success rate than the model proposed. We then choose to model it as linear because first

Table 2 Logistic regression estimates[7]

Variables in the equation	*B*	*SE*	*WALD*	*df*	*Exp(B)*
AGE	−2.686*	0.020	18,961.64	1	0.07
COMPLEX	1.733*	0.044	1524.10		5.66
IPC	0.322*	0.006	2916.67	1	1.38
CLASS#			4462.96	2	
Class (Open)	3.022*	0.293	106.44	1	20.52
Class (Closed)	4.417*	0.293	226.97	1	82.81
Constant	2.975*	0.297	100.05	1	19.59

baseline CLASS = "Other" *p < 0.01—Overall percentage correct = 80.7%. Nagelkerke R Square = 0.40—Initial –2 Log Likelihood = 108,518—Final –2 Log Likelihood = 78,772—(LR test p < 0.01)—Sample size = 19,093 words—Cases weighted by sentence length. Exp(B) = Odds Ratio (OR)

likelihood ratio (LR) test is significant, indicating that the logistic model provides a better fit to the data than the intercept-only model. A data weighting process was applied in order to take into account the length of the sentence from which the words have been extracted. As a result, with a cut-off = 0.5, overall percentage of correctly classified cases are equal to 80.7%. The Wald test in Table 2 suggests that AGE is the main regressor: as it increases, the likelihood of reporting a higher WPV decreases consistently. For every year of age, the odds of WPV decreased by roughly 90%. COMPLEX works differently: an increase in lexical richness causes an increase in WPV too, but this relation becomes weaker over time thanks to a growing mastery of his language.

As for the COMPLEX variable, we can observe how an increase in IPC determines an increase in WPV. As expected, an increase in phonetic complexity results in an increase of error rate.

This relation should become weaker over time because *Adrien* WPV globally decreases with age, as expressed by the lexical development graphs (Figs. 2, 3 and 4).

The Figs. 5, 6 and 7 plot the WPV probability profiles with respect to the variable CLASS, based on nine pre-defined scenarios and three hypothetical levels of IPC (0; 4; 8). To give some examples, a word such as "maman" has an IPC score of 0 because it is the simplest pronounceable word (for this reason it is almost universal) and the same goes for "papa" ("m" and "p" are both bilabials), while the word "table" has a score of four and the word "comprendre" a score of 10 (because of its length and

language acquisition is a highly non-linear phenomenon and the only certainty linguists have is that – roughly speaking – it develops in a cumulative way over time. We tried to model the interaction effects between COMPLEX, IPC and CLASS too, but it turned out to be less precise than the model proposed: in fact, COMPLEX showed a counterintuitive result in which its increase in value causes a decrease in WPV (models are available on request).

[7] All calculations are performed with STATA ver. 15.

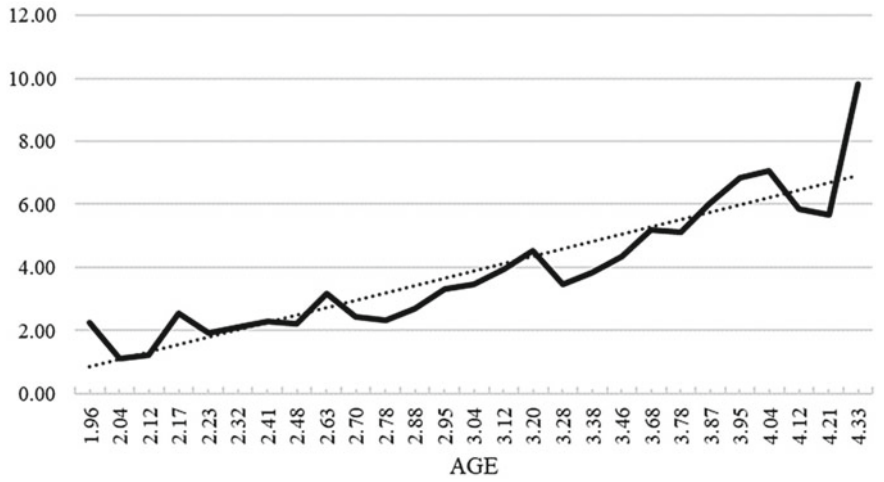


Fig. 2 Mean number of words per utterance according to AGE[8] (dotted line = trend)

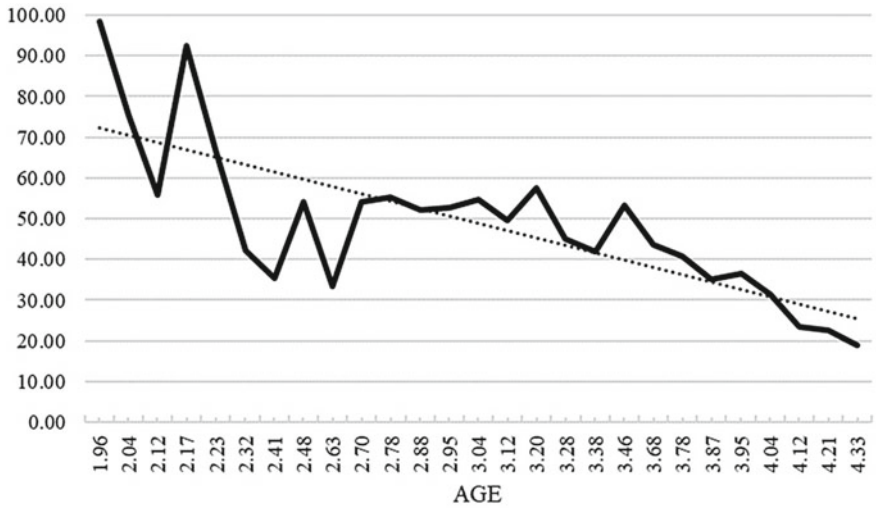


Fig. 3 Sentence phonetic variation rate (in percent)[9] by AGE (dotted line = trend)

[8] The AGE in the abscissa axis refers to the period of the video recordings (see CoLaJE database [4] for more details).

[9] The Sentence Phonetic Variation Rate (SPVR) is the ratio between the number of phonetic variations (the number of differences detected between "pho" and "mod") and the total numbers of words. SPVR can assume the value 0% when the child does not make any errors and 100% when the child does not correctly pronounce any of the words contained in the sentence [17].

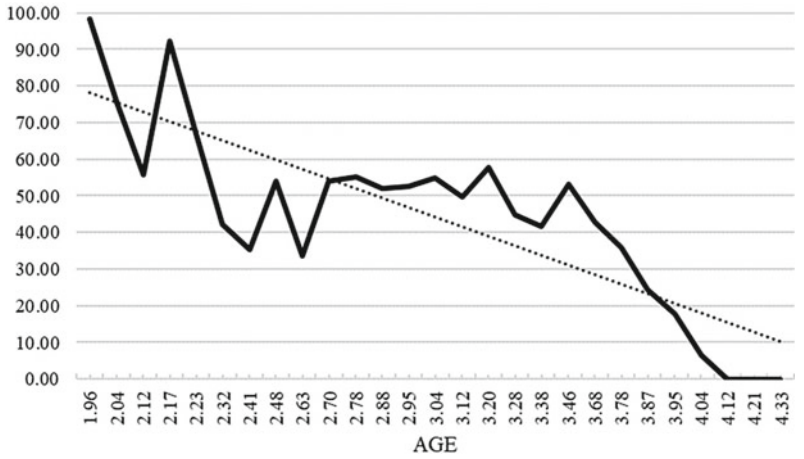


Fig. 4 Word phonetic variation (in percent) by AGE (dotted line = trend)

because of the high number of consonants).[10] Moreover, we consider a range of 1 to 5 years for the AGE variable while for the COMPLEX variable a (hypothetical) range of 0.1 to 0.9. To give an example "A1-C0.1" means one year of age, and a percentage of distinct words of 10% per sentence (where "A" stands for AGE and "C" stands for COMPLEX). As we can see from the database structure [4], ages are expressed in a "year_month_day" structure, such as *ADRIEN-6-1_09_08*, which stands for *Adrien*'s sixth sample recorded at one year, nine months and 8 days. In the three following graphs, ages are represented on the X axis and turned into a float value in order to provide a more readable line of values.

COMPLEX determines an increase in WPV because a longer and more structured utterance causes an increased cognitive load in motor planning, as already pointed out in a previous study [6].

It can be observed how OPEN class words are easier to learn compared to CLOSED class words (OPEN has 20.52 times the odds of WPV compared to OTHER while CLOSED has 82.81 times the odds of WPV compared to OTHER): the difference between the two profiles shows an (almost) constant value of 0.2 up to the age of 4 years old. When the child has almost completed his growth (after 4 years) the two profiles tend to be similar. This is because children are more at ease with naming things and persons with their names instead of using more abstract pronouns which impersonally refer to them, and because verbs are easier to put in a sentence than auxiliaries, whose place must respect precise grammar rules that require time

[10] To have an idea of what 0 or 4 or 8 mean, you can type a word in the link [10]: http://igm.univ-mlv.fr/~gambette/iPhocomp/.

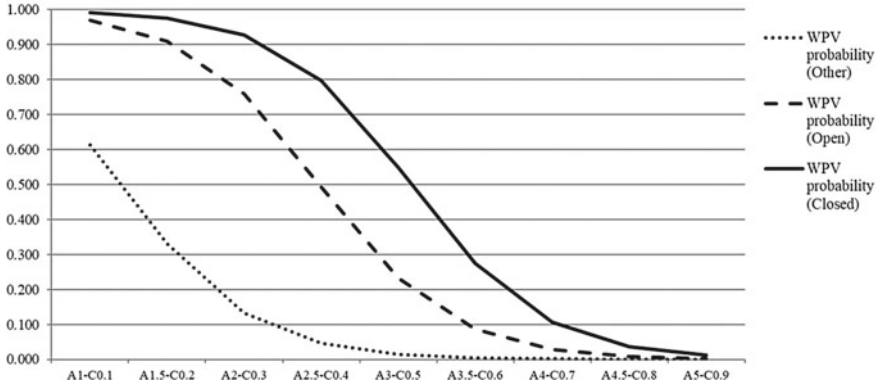


Fig. 5 Predicted probability according to 9 scenarios by CLASS category—In abscissa: A = Age (1–5 years); C = Complex Index (range 0.1–0.9)—IPC = 0

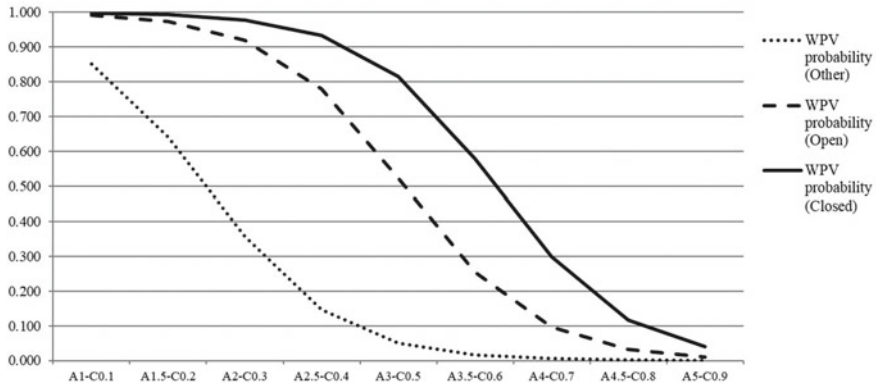


Fig. 6 Predicted probability according to 9 scenarios by CLASS category—In abscissa A = Age (1–5 years); C = Complex Index (range 0.1–0.9)—IPC = 4

to be learned.[11] We have demonstrated these two grammatical learning dynamics in a recent paper on clustering applied to first language acquisition [16].[12]

In Figs. 5, 6 and 7, AGE and COMPLEX have been fixed to see how IPC influences WPV if taken alone. We can see how the difference between OPEN and CLOSED classes remains the same in addition to an overall shift on the right side of the graph. This shift from left to right represents how words having a higher IPC will be learnt after simpler words (IPC = 0). A similar pattern can be found in the more focused graph immediately below (Fig. 8), where WPV evolution is represented

[11] The past participle form in French could be given as an example.

[12] A graphic visualization of this work can be found at this link [17]: http://advanse.lirmm.fr/EMClustering/.

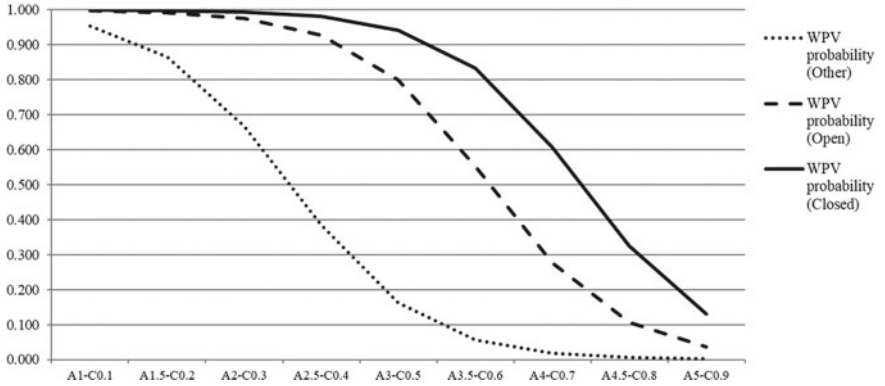


Fig. 7 Predicted probability according to 9 scenarios by CLASS category—In abscissa: A = Age (1–5 years); C = Complex Index (range 0.1–0.9)—IPC = 8

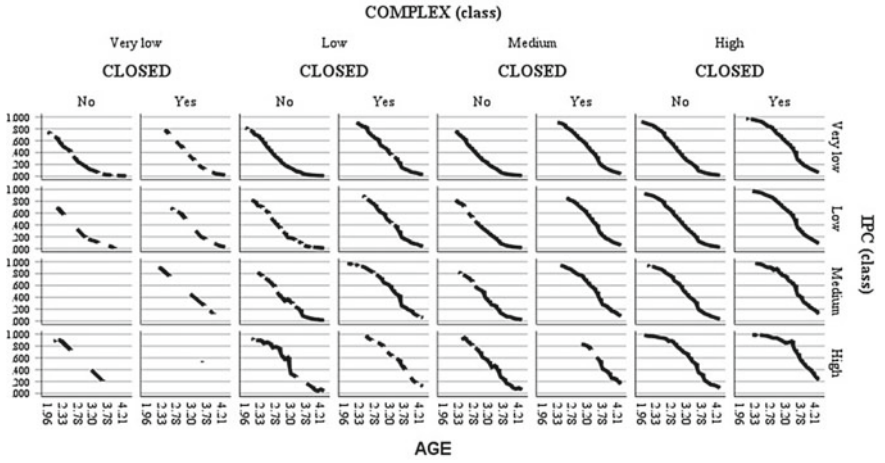


Fig. 8 Predicted probability for Eq. (1) according to AGE by COMPLEX, IPC and CLASS. (COMPLEX: Very low < = 0.25; Low: > 0.25 and <= 0.5; Medium: > 0.50 and <= 0.75; High > 0.75); (IPC: Very low = 0; Low: = 2; Medium: >2 and <= 3; High > 3); (CLASS: CLOSED (Yes) OPEN (No))

according to all the possible combinations of COMPLEX, IPC and CLASS[13] (divided into classes). Graphs to the top-left of the figure represent words contained in low type/token ratio utterances, while those to the bottom right represent words contained in high type/token ratio utterances.

[13] Considering the low number of "OTHER" type words, only the most important classes were taken: CLOSED and OPEN.

4 Conclusion and Future Directions

It is not straightforward to evaluate our results with respect to other similar studies in first language acquisition, since the corpus of interest and the language often differ. However, we think that the proposed logistic regression model could attempt to represent child language acquisition through quantitative and graphical tools. The predicted outcome is fairly good but needs to be improved by taking into account how the place a word occupies in the sentence structure influences the WPV. Attempts to create a model closer to child development in which the complexity of a sentence influences (and is influenced by) the grammatical elements contained within, turned out to be too difficult and unpredictable. According to this perspective, it seems to be better to model these regressors in the simplest possible way: by doing so, our aim becomes to exclusively focus on their main effects.

Be that as it may, this can only be true at an initial stage of research: the statistical model should be applied to other similarly sampled children. By doing so, it would become possible to test the generalizability of the claims made in this paper and improve current knowledge on first language acquisition by comparing different children, as well as comparing children learning languages of similar grammatical structures [21]. Compared to the preliminary version of this work, we succeeded in improving the overall percentage of correctly classified cases (from 72.2% to 80.7%). We are convinced that this is due to the fact that IPC gives a realistic score that needed to be counted in the multiple logistic regression model: by doing so, the set of regressors defining WPV will be closer to reality. IPC is in fact able to reflect phonetic details (at the level of perceptive and articulatory differences) that were not considered in the previous research.

However, we are planning to improve the research: it could be done by using the Index of Syntactic Complexity (ISC) as developed by Szmrecsanyi [22]. This index provides an additional way to attribute a specific weight to different grammatical elements (i.e. part of speech tags). ISC of a given context could be established by counting linguistic tokens that can be considered *telltale* signs of increased grammatical subordinateness and embeddedness [22]. In fact, some grammatical elements are more abstract than others for example, a subordinating conjunction such as “puisque” or “parce que” (corresponding to the English forms “since” and “because”) or relative pronouns such as “qui”, “que”, “duquel” (corresponding to the English forms “who”, “whose”, which”) are usually used to structure long and semantically rich sentences. We are currently working on a Python script that could operationalize this index by recognizing these specific words in the POS tags classes. We suppose that—linguistically speaking—a similar improvement could be if we add ISC to the logistic regression model. ISC would be complimentary to COMPLEX (type/token ratio) in the sense that the latter gives us a coarse-grained image of the number of distinct words per utterance (by removing words being repeated more than once) while ISC gives a fine-grained image regarding each word of the utterance by taking in account its specific grammatical role, which is subsequently weighted according to the subordinateness and embeddedness criteria. It could be fair to say that ISC

could improve the overall ratio to a value similar to that of IPC. To check the validity of our results we plan to apply the same method to another child of the CoLaJE database, to look at differences and similarities in the regression coefficients.

To date, what we have done is compare these coefficients to a deep learning-based study conducted on the same children spoken language samples. In this study, a Convolutional Neural Network [26] trained on the same samples returns the probability of a word being correctly pronounced at any given age. Although not directly comparable, it is possible to check the similarity between the WPV and the predicted output of this neural network of any given word at any given age. By doing so, we can obtain a further validation of our logistic regression model's predictions from another Natural Language Process method. A new research project on these themes is currently in progress.

References

1. Ambridge, B., Kidd, E., Rowland, C.F., Theakston, A.: The ubiquity of frequency effects in first language acquisition. J. Child Lang. **42**, 239–273. Cambridge University Press (2015)
2. Briglia, A.: Statistical and computational approaches to first language acquisition. Mining a set of French longitudinal corpora (CoLaJE). Linguistics. Université Paul Valéry Montpellier 3 (France); University of Messina (Italy) (2021)
3. Briglia, A., Mucciardi, M., Sauvage, J.: Identify the speech code through statistics: a data-driven approach, Book of Short Papers SIS. (2020)
4. CoLaJE Corpus: http://colaje.scicog.fr/index.php/corpus (2020)
5. Colombo, M., Elkin, L., Hartmann, S.: Being realist about Bayes, and the predictive processing theory of mind. Br. J. Philos. Sci. **72**(11), 185–220 (2020)
6. Didirkova, I., Dodane, C., Diwersy S.: The role of disfluencies in language acquisition and development of syntactic complexity in children. DISS 2019, Budapest, Hungary (2019)
7. Ferrer, I.C., Solé, R.V.: The small world of human language. Proc. R. Soc. Lond. B.2682261–2265 (2001)
8. Friston, K.: Life as we know it. J. R. Soc. Interface **10** (2013)
9. Hosmer, D., Lemeshow, S.: Applied logistic regression. Wiley, New York (1989)
10. Index of Phonetic Complexity. http://igm.univ-mlv.fr/~gambette/iPhocomp/ (2021)
11. Jakielski, K.: Quantifying phonetic complexity in words: an experimental index. Child Phonology Conference, Cedar Fallas, IA (2000)
12. Lee, H., Gambette, P., Barkat-Defradas, M.: iPhocomp: calcul automatique de l'indice de complexité phonétique de Jakielski. JEP 2014, XXXè édition des Journées d'Etudes sur la Parole, Le Mans, France, pp. 622–630, Actes de la XXXe édition des Journées d'Etudes sur la Parole (2014)
13. Mac Neilage, P.: The frame/content theory of evolution of speech production. Behav. Brain Sci. **21**(4), 499–511 (1998)
14. Mac Whinney, B.: The childes project: tools for analysing talk, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ (2000)
15. Morgenstern, A., Parisse. C.: The Paris corpus. French Lang Stud **22**, 7–12. Cambridge (2012)
16. Mucciardi, M., Pirrotta, G., Briglia, A.: EM Clustering method and first language acquisition. In: Book of Short Papers Models and Learning for Clustering and Classification (2021)
17. Mucciardi, M., Pirrotta, G., Briglia, A. Sallaberry, A.: Visualizing cluster of words: a graphical approach to grammar acquisition. In: Book of Abstracts and Short Papers CLADAG 2021 (2021)

18. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.J.: Stanza: a python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)
19. Saffran, J.: Statistical language learning: mechanisms and constraints. Curr. Dir. Psychol. Sci. **12**(4), 110–114 (2003)
20. Sagae, K.: Tracking child language development with neural network language models. Front. Psychol. **12**, 674402 (2021)
21. Sekali, M.: First language acquisition of French grammar (from 10 months to 4 years old). French Lang. Stud. **22**, 1–6 (2012)
22. Szmrecsanyi, B.: On operationalizing syntactic complexity. In: Purnelle, G., Fairon C., Dister A. (eds.), Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis, vol. 2. Presses Universitaires de Louvain, Louvain-la-Neuve (2004)
23. Tomasello, M., Stahl, D.: Sampling children's spontaneous speech: how much is enough? J. Child Lang. **31**, 101–121 (2004)
24. UD (Universal Dependencies).: https://universaldependencies.org (2021)
25. Yamaguchi, N.: What is a representative language sample for word and sound acquisition? Can. J. Linguist., Univ. Tor. Press. **63**(04), 667–685 (2018)
26. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision—ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer (2014)

Adaptive COVID-19 Screening of a Subpopulation

Fulvio Di Stefano and Mauro Gasparini

Abstract Methods are sought to test adaptively whether a subpopulation proportion follows the same time evolution as the population proportion. The motivating case study is the COVID-19 screening in a university community, taking into account the time evolution of the pandemic in the whole country.

Keywords Outbreak detection · Control charts · Dynamic threshold · SARS-CoV-2

1 Introduction

During the ongoing pandemic era, the need arises among members of certain working, studying or social communities to undergo a reinforced screening to immediately identify outbreaks within the community, as outlined for example in [1]. In our case study, a screening process is planned within Politecnico di Torino (POLITO), a public university, to prevent clusters among students who come in person to attend classes (most classes are also given in an online mode) during the first semester of academic year 2021–2022, running from 27 September 2021 to 14 January 2022. Based on the case study, the aim of this work is to compare methods to test, repeatedly and dynamically, whether a sub-population of interest is on average similar to the general population with respect to a certain binary characteristic (e.g. infected/not infected) with a time-changing distribution.

Outbreak detection is widely treated in literature and many studies have been conducted on the topic [2, 3]. Tukey's fences [4] and other well known static methods for outlier detection [5] are aimed at identifying excessive presence of some char-

F. Di Stefano (✉) · M. Gasparini
Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange", Politecnico di Torino, Turin, Italy
e-mail: fulvio.distefano@polito.it

M. Gasparini
e-mail: mauro.gasparini@polito.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_8

acteristic in random samples. Attribute control charts (see for example [6]) apply the same ideas to time series data. This work extends those models by introducing forecasting techniques to obtain a general methodology which combines the forecasting of the characteristic with the detection of the excess of that characteristic in a subpopulation. Thus, the main advantage of control charts, which perform well in detecting sudden and large deviation of the characteristic of interest [3], is combined with a forecasting technique which takes into account the variability of the data. The methodology is then applied to the POLITO case study to obtain an adaptively varying threshold for the COVID-19 screening in POLITO, taking into account the time evolution of the pandemic in the whole country, i.e. an alert threshold which is not fixed but is able to adapt to the predicted future evolution of the pandemic.

2 Methods

2.1 Modeling Time Evolving Proportions

Let $P_t \; \forall t \geq 0$ denote the proportion of individuals in a general population of approximately constant size N_P who carry a characteristic of interest. Suppose also that P_t is an unknown stochastic process over time, meaning that the number of individuals who carry the characteristic is a random variable with a time-changing distribution. Let $N_S \leq N_P$ be the size of a well-defined subpopulation of interest. Let p_t $\forall t \geq 0$, be proportion of individuals who carry the characteristic of interest in the subpopulation: p_t is a distinct stochastic process over time t. If the subpopulation is conformal to the general population it is a subset of, i.e. the subpopulation and the general population are homogeneous at all possible scales of observations, p_t should be approximately equal to P_t. However, the characteristic of interest may evolve differently in the subpopulation with respect to the general population. At a given time t_0, estimates of P_t for some previous time steps $t \leq t_0$ are available. These estimates, denoted as $\hat{P}_t \; \forall t \leq t_0$, are based on samples of varying sizes, but this extra source of uncertainty is ignored in this work, for reasons explained in detail later on. The following methodology aims at predicting P_{t_0+1}, having obtained estimates of P_t, $t \leq t_0$, to obtain a statistical test of whether the subpopulation proportion p_{t_0+1} is significantly larger than P_{t_0+1}. The interest in one-sided upper tests is due to the fact the we are concerned with the possibility of an excessive proportion p_t with respect to P_t, as exemplified by the POLITO case study, where evidence for an excessive proportion in the subpopulation would cause the reinforcing of restrictive measures such as confinement or distance learning.

2.2 Forecasting Using ARMA Models

Given the observed time series $\hat{P}_t \ \forall t \leq t_0$, one of the objectives is to forecast the future value P_{t_0+1}. The reason is two-fold: to have a reference value which moves over time, according to the progress of the underlying characteristic of interest in the general population, and to incorporate all the information collected up to this point in the next prediction.

Well-known methods to forecast the future values of a time series when the underlying process is unknown are ARMA models (a good textbook introduction can be found in [7]), very popular in the econometric literature. ARMA models have been generalized in several ways and can easily be adapted to many time series. Far from claiming ARMA models are sufficient for all predictions, we propose instead to use them as a working tool to update a population reference value.

Each ARMA model is constructed to have two order parameters, denoted (p, q), which have to be "identified" based on data with an empirical model selection procedure, and several unknown regression parameters, which have to be estimated based on data. A generic ARMA(p, q) model can be applied to our variable of interest P_t, after a preliminary logarithmic transformation, to account for the inherent positiveness of P_t (the fact that P_t is also bounded above by 1 can be neglected, since P_t is usually a small value, way closer to 0 than to 1):

$$\log(P_t) = K + a_1 \log(P_{t-1}) + ... + a_p \log(P_{t-p}) + \epsilon_t + b_1 \epsilon_{t-1} + ... + b_q \epsilon_{t-q}$$

where K is the underlying mean of the process, the ϵ_t are the error terms, $a_i \ i = 1, ..., p$ are the coefficients associated to the auto-regressive part of the model and b_j $j = 1, ..., q$ are the coefficients associated to the moving average part of the model. The ϵ_t are assumed to be independent and identically distributed, following a normal distribution $N(0, \sigma^2)$.

The estimation of the order and the coefficients of the model is widely treated in the literature [7, 8] and goes beyond the scope of this work. However, it can briefly be mentioned that the values of p and q can be chosen to minimise the Bayesian information criterion (BIC) [8, 9].

In our setup, the estimates $\hat{P}_t \ \forall t \leq t_0$ can be used as surrogate P_t to estimate the parameters of the model. Therefore, having determined the order (p, q) of the model and estimated the parameters $\hat{a}_i$ and $\hat{b}_j$, for $i = 1, ..., p$, $j = 1, ..., q$, $\hat{K}$ and $\hat{\sigma}^2$ and having obtained the realizations of the errors $\hat{\epsilon}_t \ \forall t \leq t_0$, we can proceed with the forecasting. Given all available information up to t_0, the prediction for the next time-step can be calculated as:

$$\log(\tilde{P}_{t_0+1}) = \hat{K} + \hat{a}_1 \log(\hat{P}_{t_0}) + ... + \hat{a}_p \log(\hat{P}_{t_0-p}) + \hat{b}_1 \hat{\epsilon}_{t_0} + ... + \hat{b}_q \hat{\epsilon}_{t_0-q}$$

This quantity can be thought to be approximately normally distributed with variance σ^2, estimated by $\hat{\sigma}^2$ for practical purposes.

2.3 Detecting Excessive Presence of the Characteristic of Interest in the Subpopulation

Given the forecast for the next time period in the whole population, inference can be made over the presence, and in particular over possible excessive presence, of the characteristic of interest in the subpopulation. Suppose then a random sample of n_S out of N_S individuals are tested for the presence of the characteristic in the subpopulation. Let X_t be the number of individuals who carry the characteristic of interest among the tested individuals n_S at time t; then X_t has a hypergeometric distribution with (discrete) density

$$\text{Prob}_{p_t}(X_t = x) = \frac{\binom{N_S p_t}{x}\binom{N_S(1-p_t)}{n_S - x}}{\binom{N_S}{n_S}}$$

which, for N_S large compared to n_S, can be approximated by the binomial density

$$\text{Prob}_{p_t}(X_t = x) = \binom{n_S}{x} p_t^x (1 - p_t)^{n_S - x}$$

where p_t is an unknown parameter evolving over time. We would like to test formally, at level $1 - \alpha$, the system of hypotheses

$$\begin{cases} H_0 : p_{t_0+1} = P_{t_0+1} \\ H_A : p_{t_0+1} > P_{t_0+1} \end{cases}$$

Following the methods described in the previous section, at time t_0 we have a forecast $\tilde{P}_{t_0+1}$ for the next period, accompanied by an estimate $\hat{\sigma}$ of its uncertainty. We can proceed in different ways.

Method 1: direct thresholding. Use the normal approximation retrieved from the ARMA model to obtain an explicit threshold τ_{1,t_0+1} and use the decision rule

"Reject H_0 if $X_{t_0+1} > \tau_{1,t_0+1} := n_S \exp(\log(\tilde{P}_{t_0+1}) + z_{1-\alpha}\hat{\sigma})$",

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the normal distribution.

Method 2: binomial testing. A more traditional approach would be to perform a standard binomial test at a significance level approximately equal to $1 - \alpha$ (due to the discreteness of the binomial distribution), with decision rule

"Reject H_0 if $X_{t_0+1} > \tau_{2,t_0+1}$",

where τ_{2,t_0+1} is the $(1 - \alpha)$-quantile of the binomial distribution with parameters n_S and $\tilde{P}_{t_0+1}$. This method disregards the uncertainty about P_{t_0+1}.

Method 3: normal testing. If n_S is large enough, the binomial distribution can be approximated by a normal distribution, so that the following decision rule is obtained:

"Reject H_0 if $X_{t_0+1} > \tau_{3,t_0+1} = n_S \tilde{P}_{t_0+1} + z_{1-\alpha}\sqrt{n_S \tilde{P}_{t_0+1}(1 - \tilde{P}_{t_0+1})}$".

The idea behind these techniques is to combine the main advantage of control charts, which perform well in detecting sudden and large deviation from the average [3], with a forecasting technique which is able to take into account the evolution of the characteristic of interest over time.

2.4 A Naive Fixed Threshold

The three methods previously presented are opposed to a naive method which consists in giving an alert if, at some α level type I error, the null hypothesis $H_0 : p_t = p_N$ at time $t \geq 0$ is rejected, using binomial testing. p_N is a pre-determined level of diffusion of the characteristic of interest in the subpopulation which is considered acceptable.

Method 4: naive fixed threshold binomial testing. Binomial testing using a fixed threshold results in the following decision rule.

"Reject H_0 if $X_{t_0+1} > \tau_N$",

where τ_N, is the $(1 - \alpha)$-quantile of the binomial distribution parameters n_S and p_N.

This procedure is formally equivalent to a binary attribute control chart [6] and does not take into account the evolution of the characteristic of interest over time.

3 A Case Study: COVID-19 Testing at Politecnico di Torino

3.1 Current Testing

In POLITO, during the first semester of the academic year 2021–2022, an estimate of $N_S = 21870$ students plan to attend classes in person, and the university screens sample of students on site to develop a system of alert to predict possible outbreaks. This cluster detection system works alongside the screening and the prevention measures every national system is implementing. In particular, $n_S = 250$ oropharyngeal swabs are carried out every Monday, Wednesday and Friday, due to procedural and technical constrains, leading to the total of 750 swabs weekly. We will discuss the previous test using $n_S = 250$, whereas in practice the same test will be repeated three times a week. The test will be conducted on a random sample of students who have booked to attend their lessons on that particular day. Students can refuse to be tested. However, up to the 3 December 2021, the number of refusals to the tests has been of a few units. Usually, students who are symptomatic and acknowledge to be positive do not come in presence, given the possibility to follow online lessons. This

fact remarks the necessity of a screening procedure which is able to capture possible unnoticed clusters, like the ones caused by asymptomatic cases, inside the university.

Following Sect. 2, the probability of a positive test in a particular day $t \geq 0$ is denoted by p_t. Therefore, following Method 4, an alert will be given if, at some α level type I error, the null hypothesis $H_0 : p_t = p_N$ at time $t \geq 0$ is rejected. Using $\alpha = 0.20$, $n_S = 250$ and $p_N = 0.015$, a rough guess of the average pandemic situation in the country in September 2021, we obtain $\tau_N = 5$ (the reason for using such a large level of type I error is explained below). This procedure is the one currently implemented in POLITO, which the authors have helped to set up. Up to the 3 December 2021, the screening has produced very few positive tests, far below the threshold, due to particularly stringent rules for accessing the POLITO site.

The current procedure does not take into account the fact that the pandemic is evolving over time. In particular, in Italy the evolution of the pandemic is monitored by Istituto Superiore di Sanità and Protezione Civile, who provide day by day data on the evolution of the pandemic [10]. These data give a clue on how the pandemic is evolving in Italy and, if the pandemic is worsening in the whole country, it is expected that it will also worsen in POLITO, provided that the above homogeneity assumption holds. The use of national data and not regional (Piemonte) data is twofold. Notwithstanding the fact that the current regulation still allocates Italian regions in different risk areas according to the regional spread of the pandemic, all restrictions of movements (like travelling between regions) and many social restrictions do not apply to people who have obtained the so-called "green pass certification", which is mandatory also to physically access universities. Moreover, just over a third of POLITO students residing in Italy come from the Piemonte region: a huge number of students come from the south of Italy and a big component comes also from the centre and north-east of Italy. Therefore, given the heterogeneous regional composition of the students which plan to attend class in person and the freedom guaranteed to them by the current regulation, which permits to people in possession of the green pass certification to not experience any of the regional constraints in place, national data are believed to better represent the considered subpopulation.

In Fig. 1, the weekly percentage of positive tests in Italy since the beginning of the pandemic is shown. Now, for reasons related to the way these data are collected in Italy, they do not properly represent the proportion of infected people at time t. As a matter of fact, molecular tests are much more precise than antigenic tests and are carried out mainly to confirm cases reported via the latter ones, increasing the estimate of positive cases since they may be reported multiple times. However, the data on antigenic tests decrease the estimate of positive cases, since a huge number of these tests are carried out by not vaccinated people to access many working activities, in compliance with the current regulation. Moreover, the number of weekly antigenic tests is approximately twice the number of molecular tests in this phase of the pandemic, but it is not consistent over time. For these reasons, the percentage of positive tests probably overestimates the true proportion of infected people at time t, However, it is difficult to quantify how large the overestimation is, but the expected percentage of positive tests in POLITO might be lower than those numbers. Rather than arbitrarly reduce by a significant fraction those percentages, we prefer to use

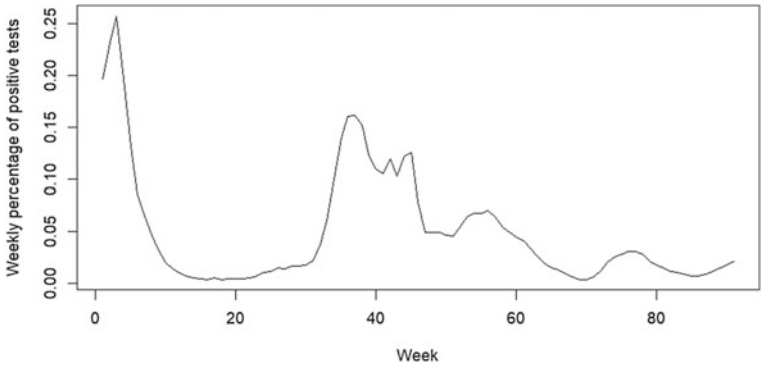


Fig. 1 Percentage of weekly positive SARS-CoV-2 tests in Italy since the beginning of the pandemic (Week 0 is 24-02-2020) up to the end of November 2021 (Week 91 is 28-11-2021)

the original data but work with a large type I error, following the criterion that in this situation prevention is better than cure, and a false alarm does not entail dramatic consequences. In particular, the authors will help in analysing the data coming from the screening process and, if necessary, start an alarm. Given the alarm, a series of containment measures will be taken, starting from retesting the positive people with molecular tests—more precise than the rapid tests used in POLITO—then escalating to quarantineing people and possibly switch to online teaching for entire classes. We will therefore use a large $\alpha = 0.20$ and take as our $\hat{P}_t \ \forall t \geq 0$ the values shown in Fig. 1. These are calculated by dividing the total number of new cases reported in a particular week over the total number of tests (both antigen and molecular) conducted in that particular week. The rationale for using percentages instead of counts is that they are known to help in facing inconsistency in reporting of cases[2], a well-known topic in outbreak detection.

Using the dataset from the Istituto Superiore di Sanitá and Protezione civile and the methodology described earlier, a forecast on the percentage of positive tests in Italy next week can be obtained. If the pandemic is worsening (or improving) in the whole country, the ARMA(p, q) model will provide an adequate prediction, which can be used for a more accurate alarm system with respect to a fixed threshold. The use of the weekly time series permits to have a quantity which can be compared to the percentage of positive tests among the students at POLITO. Moreover, the data are grouped weekly for two reasons: the number of tests during the week is not constant but depends on the day of the week, and the fact that the tests at the Politecnico are not planned daily, as for those in the whole country.

Finally, due to the highly non-stationary evolution of the pandemic, which depends on many factors such as restrictive measures, vaccines, seasonality, heterogeneous intensity of screening (e.g. differing search rates for asymptomatic individuals), it is recommendable to use only the last portion of the time series; in this work we use only the last 16 weeks of observations and disregard the previous ones dinamically in case we are interested in different t_0 times.

More accurate methods to predict the number of SARS-CoV-2 positives in a general population are available in literature: the SIS model [11], the SIPRO model [12], the SIDARTHE model [13] and its extentions [14], the Covasim model [15], and many others (for example [16–18]). However, the number of weekly tests in Italy is not constant over time and social and economical measures are continuously taken by the governments to contain the pandemic, making a precise estimation of P_t very hard. In this situation, far from proposing ARMA models as a well-thought realistic model for epidemic prevalence, the proposal in this work is to use ARMA models as a working and adaptive tool which performs well enough to give one-time step ahead predictions, and no further in time.

3.2 A Proposal for Adaptive Testing

Using R version 4.1.2 and the package forecast [19], the methodology of Sect. 2 can be applied to the dataset. Suppose to start the estimation at week 74 after the beginning of the pandemic, i.e. 25 July 2021. At this date, the weekly percentage of positive tests in Italy is $\hat{P}_{74} = 0.025$, and the curve is slightly increasing. Using R, it is possible to fit an ARMA(p, q) model to the dataset using the previous 16 weeks' data.

The best fitting model, according to the minimization of the BIC, is ARMA(4,0). The following parameters are estimated: $\hat{a}_1 = 2.723$, $\hat{a}_2 = -3.334$, $\hat{a}_3 = 2.235$, $\hat{a}_4 = -0.705$, $\hat{K} = -3.876$, $\hat{\sigma}^2 = 0.009$. Using this model, we can forecast $\tilde{P}_{75} = 0.026$. This number is bigger than $\hat{P}_{74}$, as expected since the pandemic is slightly worsening in this period. Also, the three thresholds described in Sect. 2.3 are calculated: $\tau_{1,75} = 8$, $\tau_{2,75} = 9$, $\tau_{3,75} = 9$, where $\tau_{1,75}$ and $\tau_{3,75}$ have been rounded to the next integer. Therefore, if the number of positives on Monday, Wednesday or Friday of week 75 among the 250 swabs at the Politecnico is greater or equal than the threshold of our interest, we start an alert since we reject the null hypothesis that the proportion of positives in the university p_{75} is equal to the proportion of positives in the whole country P_{75}: there is evidence for an ongoing outbreak in POLITO.

After collecting the national data for week 75, we can proceed to estimate the threshold for the next week. And then repeat this procedure over time. Figure 2 shows the three different thresholds calculated from week 75 to week 92 since the beginning of the pandemic, using all the data available from the previous 16 weeks. Comparing this Figure with Fig. 1, it is clear that the progress of the pandemic has been adequately incorporated in the model. The decrease in the weekly percentage of positive after week 77 is captured by the variable thresholds, as it is for the uprising trend after week 86. Of the three different thresholds, τ_1 is the most conservative, resulting in the lowest number of positives to be achieved to start an alarm. On the other hand, τ_2 and τ_3 give very similar results, with τ_2 being the most conservative of the two. This is because the rationale behind the two thresholds is the same, but τ_3 is derived from a normal approximation of binomial distribution, which traces the widely used asymptotic test for proportions, while τ_2 is the exact quantile of a

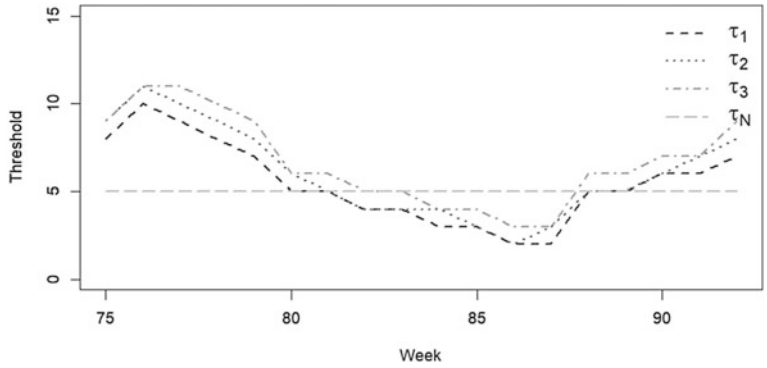


Fig. 2 The variable and fixed thresholds for the SARS-CoV-2 swabs at Politecnico di Torino

binomial distribution. In any case, all these methods outperform the fixed threshold τ_N, which is not able to capture any of the fluctuation of the progress of the pandemic: it can result in a lower threshold compared to the others when the pandemic is in an expansion phase, or in a higher threshold comparing to the others when the pandemic is in a regressive phase.

3.3 Operating Characteristics of Adaptive Testing

Given the available data on the progress of the pandemic $\hat{P}_t$, some operating characteristics can be calculated for the different thresholds to check their properties. As an example, it is possible to start from the previously calculated $\tau_{1,75} = 8$, $\tau_{2,75} = 9$, $\tau_{3,75} = 9$, $\tau_{N,75} = 5$ and the real percentage of positives tested in Italy in week 75 $\hat{P}_{75} = 0.028$, to retrieve power and type I error of the thresholds for this week. In particular, supposing that the positive tests at POLITO follow a binomial distribution with parameters n_S and p_{75}, the type I error can be defined as:

$$\alpha_{i,75} = 1 - \sum_{x=1}^{\tau_{i,75}} \binom{n_S}{x} (\hat{P}_{75})^x (1 - \hat{P}_{75})^{n_S - x}$$

for $i = 1, 2, 3, N$; while the power can be defined as:

$$(1 - \beta_{i,75}) = 1 - \sum_{x=1}^{\tau_{i,75}} \binom{n_S}{x} (3 \cdot \hat{P}_{75})^x (1 - 3 \cdot \hat{P}_{75})^{n_S - x}$$

for $i = 1, 2, 3, N$. This type I error identifies the probability to overcome the given threshold when in truth $p_{75} = \hat{P}_{75}$, resulting a false alarm. On the other hand, the

power is the probability to overcome the given threshold when in truth $p_{75} = 3 \cdot \hat{P}_{75}$, resulting a right alarm. For week 75 $\alpha_{1,75} = 0.271$, $\alpha_{2,75} = 0.168$, $\alpha_{3,75} = 0.168$, $\alpha_{N,75} = 0.705$, $(1 - \beta_{1,75}) = 0.999$, $(1 - \beta_{2,75}) = 0.998$, $(1 - \beta_{3,75}) = 0.998$, $(1 - \beta_{N,75}) = 1$.

In Figs. 3 and 4 the type I error and power for the different thresholds are presented. These are calculated using the available weekly data of $\hat{P}_t$ from week 75 to 91, as shown above. It can be seen that the fixed threshold gives an extremely high type I error but also a very high power when the pandemic is worsening, but its type I error is controlled under 0.20 and its power decreases down to 0.429 when the pandemic is in a regressive phase. This means that in a expansive phase of the pandemic there is a high risk of a false alarm, because of the more plausible high number of positives, while in a regressive phase it is realistic to not detect a possible cluster inside the university, as a consequence of the high threshold. Instead, the type I error of τ_1 is almost everywhere above 0.07 and below 0.30, with a peak of 0.431 at week 87 (when the weekly percentage of positive tests changes its convexity and starts

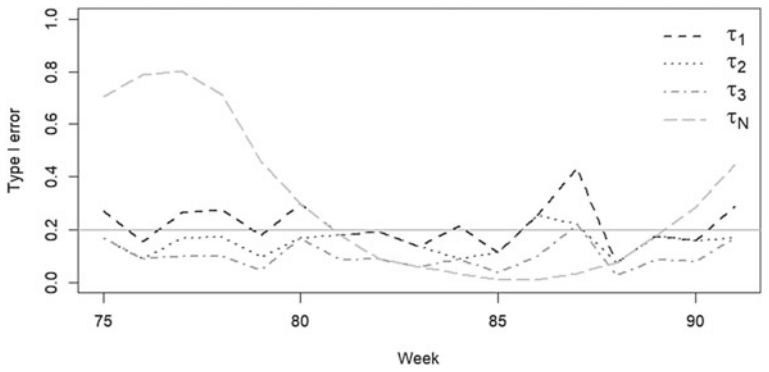


Fig. 3 Type I error for the different thresholds. The solid line is 0.2

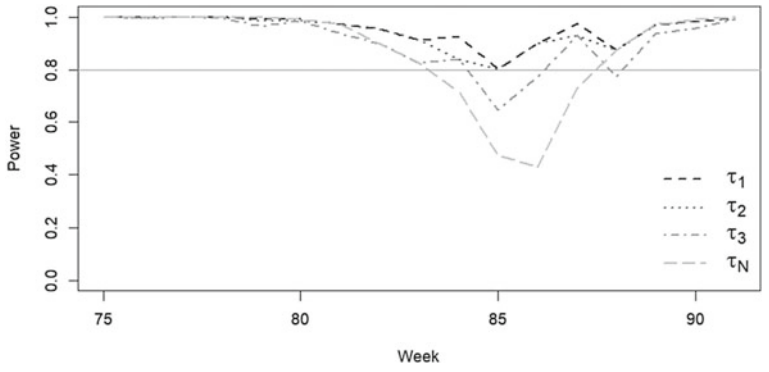


Fig. 4 Power for the different thresholds. The solid line is 0.8

Table 1 Summary of the operating characteristics of the different thresholds

Threshold	Fixed	Type I error	Power
τ_1	No	(0.07, 0.44)	(0.80, 1)
τ_2	No	(0.07, 0.26)	(0.80, 1)
τ_3	No	(0.02, 0.23)	(0.64, 1)
τ_N	Yes	(0.01, 0.80)	(0.42, 1)

growing again); however, its power is above 0.8 in all weeks. As regards τ_2 and τ_3, these give very similar results. Their type I error is controlled under 0.20 everywhere except around week 87, with τ_2 peaking at 0.256 at week 86 and at 0.221 at week 87 and τ_3 peaking at 0.221 at week 87. The power of τ_2 is above 0.8 all of the time, while the power of τ_3 is above 0.8 most of the times, except on week 86 and 88 where it goes down to 0.772 and 0.776, respectively.

A brief summary of the operating characteristics is shown in Table 1, where it can be seen that the variable thresholds have better operating characteristics with respect to a fixed threshold: the proposed methodology results in a lower number of false alarms and in improved detection of possible outbreaks.

4 Discussion

In this work, a methodology is proposed to identify how conformal a subpopulation is to a general population with respect to the distribution of a binary variable. This study was motivated by a case study on the SARS-CoV-2 tests in POLITO, to identify outbreaks inside the university via the screening process organized with oropharyngeal swabs three days a week.

Making use of a very general ARMA(p, q) model, three thresholds which vary over time have been determined. These thresholds are used to test the equality of the proportion of individuals with the characteristic of interest in the subpopulation and the general population. Via the case study, it has been shown that the three presented variable thresholds are able to capture the progress of the underlying process, outperforming a fixed threshold in terms of operating characteristics. The three presented thresholds exhibit different properties: threshold τ_1 performs very well in terms of power, but type I error is not controlled at the however large $\alpha = 0.20$ level we decided to work with; thresholds τ_2 and τ_3 have controlled type I error at a $\alpha = 0.20$ level, but a little less power with respect to τ_1.

Some limitations and future extension of this work regard the possibility to use a more accurate model to predict the COVID-19 pandemic in Italy and therefore obtain more accurate thresholds for the case study. However, this work aims at being very general, in order to be adapted to various possible scenarios.

Acknowledgements The authors would like to thank Paola Lerario and Maurizio Galetto for insights and details on the organization of the screening procedure in Politecnico di Torino and Enrico Bibbona for some critical discussions. The authors would like to thank also two anonymous reviewers, who greatly helped to improve the quality of the manuscript.

References

1. European Center for Disease prevention and Control: COVID-19 clusters and outbreaks in occupational settings in the EU/EEA and the UK (2020)
2. Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R., Moore, A.W.: Algorithms for rapid outbreak detection: a research synthesis. J. Biomed. Inform. (2005). https://doi.org/10.1016/j.jbi.2004.11.007
3. Leclère, B., Buckeridge, D.L., Boëlle, P.Y., Astagneau, P., Lepelletier, D.: Automated detection of hospital outbreaks: a systematic review of methods. PLOS ONE (2017). https://doi.org/10.1371/journal.pone.0176438
4. Tukey, J.: Exploratory Data Analysis. Addison-Wesley Pub, Co (1977)
5. Hawkins, D.M.: Identification of Outliers. Springer (1980)
6. Montgomery, D.C.: Introduction to Statistical Quality Control. Wiley (2019)
7. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. Springer (2009)
8. Stoica, P., Selen, Y.: Model-order selection. IEEE Sig. Process. Mag. (2004). https://doi.org/10.1109/msp.2004.1311138
9. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. (1978). https://doi.org/10.1214/aos/1176344136
10. https://github.com/pcm-dpc/COVID-19/. Cited 20 Dec 2021
11. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. In: Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1927). https://doi.org/10.1098/rspa.1927.0118
12. Amongero, M., Bibbona, E., Mastrantonio, G.: Analysing the Covid-19 pandemic in Italy with the SIPRO model. Book of short papers SIS 2021
13. Giordano, G., Blanchini, F., Bruno, R., et al.: Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat. Med. (2020). https://doi.org/10.1038/s41591-020-0883-7
14. Giordano, G., Colaneri, M., Filippo, A.D., et al.: Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. Nat. Med. (2021). https://doi.org/10.1038/s41591-021-01334-5
15. Kerr, C.C., Stuart, R.M., Mistry, D., et al.: Covasim: an agent-based model of COVID-19 dynamics and interventions. PLOS Comput. Biol. (2021). https://doi.org/10.1371/journal.pcbi.1009149
16. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G.: An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom. J. (2020). https://doi.org/10.1002/bimj.202000189
17. Fokas, A.S., Dikaios, N., Kastis, G.A.: Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. J. R. Soc. Interface (2020). https://doi.org/10.1098/rsif.2020.0494
18. Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M.: Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science (2020). https://doi.org/10.1126/science.abb5793
19. https://cran.r-project.org/web/packages/forecast/index.html. Cited 20 Dec 2021

Alternative Probability Weighting Functions in Behavioral Portfolio Selection

Diana Barro, Marco Corazza, and Martina Nardon

Abstract We propose some portfolio selection models based on Cumulative Prospect Theory. In particular, we consider alternative probability weighting functions in order to model probability distortion. The resulting mathematical programming problem turns out to be highly non-linear and non-differentiable. So, we adopt a solution approach based on the metaheuristic Particle Swarm Optimization. We select the portfolios under the behavioral approach and perform an application to the European equity market as represented by the STOXX Europe 600 Index and compare their performances.

Keywords Behavioral finance · Portfolio selection · Cumulative Prospect Theory · Probability distortion function · Particle Swarm Optimization

1 Introduction

The literature on portfolio selection models is wide and since the founding work of Markowitz [15] has grown rapidly and has been extended along many different directions. Among them, modeling risk measures and performance evaluation are certainly crucial themes.

We apply Cumulative Prospect Theory (CPT) of [20] to the portfolio selection problem, similarly to what done in [5], with the aim of modeling optimal decisions tailored to individual attitudes to risk and loss aversion. Previously, Shefrin and Statman [19] propose a behavioral portfolio model under PT. CPT provides a framework

D. Barro · M. Corazza (✉) · M. Nardon
Department of Economics, Ca' Foscari University of Venice, Sestiere Cannaregio 873, 30121 Venezia, Italy
e-mail: corazza@unive.it

D. Barro
e-mail: d.barro@unive.it

M. Nardon
e-mail: mnardon@unive.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_9

to effectively represent a wider range of risk attitudes allowing for a larger flexibility in the description of the portfolio problem.

In the present work, we focus on the effects on the portfolio choices and performances of alternative probability distortion functions. In particular, first, we collect preliminary evidence on the role of these specifications on the investment choices. Then, we model and apply the portfolio selection problems defined under CPT. Lastly, we compare compositions and performances of the so selected Behavioral Portfolios (BP). In order to solve BP optimization problem, which is highly nonlinear and non-differentiable, we resort to an evolutionary metaheuristic, Particle Swarm Optimization (PSO).

The remainder of this paper is organized as follows. Sections 2 and 3 synthesize the main features of Prospect Theory and the probability weighting functions. Section 4 presents the BP selection models. Section 5 briefly describes the PSO. In Sect. 6, an application to the European equity market is discussed. Section 7 concludes.

2 Prospect Theory

Kahneman and Tversky [13] proposed Prospect Theory (PT)[1] as an alternative to Expected Utility Theory (EU) in order to explain actual behaviors of decision makers. In PT, individuals do not always take their decisions consistently with the maximization of EU: results related to investment decisions are evaluated in terms of potential gains and losses, defined with respect to some *reference point* (the *status quo*), instead of final wealth. Decision makers display risk aversion with respect to gains and are risk seeking with respect to losses; moreover, they are more sensitive to losses than gains of same magnitude (loss aversion). Investment choices are evaluated through a *value function*, v, which replaces the utility function.

The value function is typically concave for gains (risk aversion) and convex (risk proneness) and steeper for losses. A function with these characteristics, which is largely used in the literature (and also applied in this work), is

$$v(z) = \begin{cases} v^+(z) = z^a & z \geq 0 \\ v^-(z) = -\lambda(-z)^b & z < 0, \end{cases} \tag{1}$$

with positive parameters that control risk attitude, $0 < a \leq 1$ and $0 < b \leq 1$, and loss aversion, $\lambda \geq 1$. Function (1) is continuous, strictly increasing, and has 0 as reference point.[2]

[1] Wakker [21] provides a thorough treatment on PT.

[2] Tversky and Kahneman [20] estimated these parameters: $a = b = 0.88$, and $\lambda = 2.25$. We will refer to this set of parameters as *TK sentiment*.

In PT, decision makers have a biased perception of probabilities of outcomes: medium and high probabilities tend to be underweighted and low probabilities of extreme outcomes are overweighted. For each result, objective probabilities (p_i) are replaced by *decision weights* π_i, which are distorted probabilities, computed through a *probability weighting function* (or *probability distortion*) w.

In CPT [20], the prospect value depends also on the *rank* of the outcomes and the decision weights are defined as differences in transformed counter-cumulative probabilities of gains and cumulative probabilities of losses. Formally,

$$\pi_i = \begin{cases} w^-(p_{-m}) & i = -m \\ w^-\left(\sum_{j=-m}^{i} p_j\right) - w^-\left(\sum_{j=-m}^{i-1} p_j\right) & i = -m+1, \ldots, -1 \\ w^+\left(\sum_{j=i}^{n} p_j\right) - w^+\left(\sum_{j=i+1}^{n} p_j\right) & i = 0, \ldots, n-1 \\ w^+(p_n) & i = n, \end{cases} \tag{2}$$

where w^- and w^+ denote the weighting function for probabilities of losses and gains, respectively.

Hence, risk attitude and loss aversion are modeled through the value function v, whereas the probability weighting function w models probabilistic risk perception via a distortion of probabilities of ranked outcomes. Actual investment decisions depend on the shapes and interaction of these two functions.

Within this framework, the objective of a prospect investor (PI) is then the maximization of the following Prospect Value:

$$V = \sum_{i=-m}^{n} \pi_i \cdot v(z_i), \tag{3}$$

where z_i denotes negative outcomes for $-m \leq i < 0$ and positive outcomes for $0 < i \leq n$, with $z_i \leq z_j$ for $i < j$. Results are interpreted as deviations from the reference point, r_0.

Several parametric forms for the probability weighting function have been used in many theoretical and empirical studies. Some forms are derived axiomatically or are based on psychological factors. Single parameter and two (or more) parameter weighting functions have been suggested; some functions have linear, polynomial or other forms, and there is also some interest for discontinuous weighting functions. Two commonly applied probability weighting functions are those proposed by Tversky and Kahneman in [20], and by Prelec in [17]. As, in the present work, we focus on the effects on the portfolio choices and performances of the probability distortion, the next section will be devoted to a review on probability weighting functions proposed in the literature.

3 The Probability Weighting Function

A probability weighting (or probability distortion) function w is a strictly increasing function which maps the probability interval $[0, 1]$ into $[0, 1]$, with $w(0) = 0$ and $w(1) = 1$. Here we assume continuity of w on $[0, 1]$, even though in the literature discontinuous weighting functions are also considered.

Empirical evidence suggests a typical *inverse-S shape*: the function is initially concave (probabilistic risk seeking or *optimism*) for probabilities in the interval $(0, p^*)$, and then convex (probabilistic risk aversion or *pessimism*) in the interval $(p^*, 1)$, for a certain value of p^*. The *curvature* of the weighting function is related to the risk attitude toward probabilities; a linear weighting function describes probabilistic risk neutrality or objective sensitivity towards probabilities, which characterizes EU. Moreover, individuals are more sensitive to changes in the probability of extreme outcomes than mid outcomes (extreme sensitivity): small probabilities of extreme events are overweighted, $w(p) > p$, whereas medium and high probabilities are underweighted, $w(p) < p$. Empirical findings[3] indicate that the intersection, or *elevation*, between the weighting function and the 45° line, $w(p) = p$, is for p in an interval around $1/3$.

The sensitivity toward probability is increased[4] if $w(p)/p > 1$, for $p \in (0, \delta)$, and $(1 - w(p))/(1 - p) > 1$, for $p \in (1 - \epsilon, 1)$, whereas a weighting function exhibits decreased sensitivity if $w(p)/p < 1$, for $p \in (0, \delta)$, and $(1 - w(p))/(1 - p) < 1$, for $p \in (1 - \epsilon, 1)$, for some arbitrary small $\delta > 0$ and $\epsilon > 0$. Some weighting functions (e.g., the functions suggested by Goldstein and Einhorn [11]; Tversky and Kahneman [20]; Prelec [17]) display *extreme sensitivity*, in the sense that $w(p)/p$ and $(1 - w(p))/(1 - p)$ are unbounded as p tends to 0 and 1, respectively.

An inverse-S shape of the probability weighting function combines the increased sensitivity with concavity for small probabilities and convexity for medium and large probabilities. In particular, such a form captures the fact that individuals are extremely sensitive to changes in (cumulative) probabilities which approach to 0 and 1. Abdellaoui et al. [3] discuss how optimism and pessimism are possible sources of increased sensitivity.

Different parametric forms for the weighting function with the above mentioned features have been proposed in the literature, and their parameters have been estimated in many studies.[5] We report here below the functional form of some commonly applied probability weighting functions.

Karmarkar [14] considers the following function

$$w(p) = \frac{p^\gamma}{p^\gamma + (1 - p)^\gamma}, \tag{4}$$

[3] See e.g., [1, 2, 6, 7, 17].

[4] See [3].

[5] See [16] for a review.

with $\gamma > 0$. Function (4) is a special case of the two parameter family proposed by Wu and Gonzalez [22] (when $\delta = 1$):

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^\delta}, \tag{5}$$

with δ and γ positive.

Allais [4] suggests the form

$$w(p) = \frac{m'p}{a-p} \frac{[1+(m'-1)a](a-p) - a(a-1)(m\,m'-1)(1-p)}{[1+(m'-1)a][1+(m'-1)p] - (a-1)(m\,m'-1)(1-p)}, \tag{6}$$

with $w(0) = 0$, $w(1) = 1$, $m' = \frac{\partial w}{\partial p}\Big|_{p=0}$ and $m = \frac{\partial w}{\partial p}\Big|_{p=1}$. The function depends on three parameters: m, m', and a. Based on observed data (subjective answer to questionnaires) and the properties of function (6), the parameters are such that $m > 1$, $0 < m' < m, m\,m' > 1, 1 < a < m/(m-1)$. Note that parameters m and m' govern curvature of the weighting function. In particular, the author points out that m can be interpreted as an indicator of the *preference for security* and m' as an indicator of *preference for risk* for small probabilities. As $\partial w/\partial a < 0$, the parameter a can be viewed as an indicator of the preference for security given the values of m and m'; hence, it controls elevation.

A *linear in log odds* function has been proposed by Goldstein and Einhorn [11],

$$w(p) = \frac{\delta\, p^\gamma}{\delta\, p^\gamma + (1-p)^\gamma}. \tag{7}$$

The weighting function proposed by Karmarkar [14] is a special case of (7) with $\delta = 1$.

Tversky and Kahneman [20] use the Quiggin [18] functional of the form[6]

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}, \tag{8}$$

with $w(0) = 0$ and $w(1) = 1$, and $\gamma > 0$ (with some constraint in order to have an increasing function). The parameter γ captures the degree of sensitivity toward changes in probabilities from impossibility ($p = 0$) to certainty ($p = 1$). When $\gamma < 1$, one obtains the typical inverse-S shape; the lower the parameter, the higher is the curvature of the function.

Prelec [17] suggests a two parameter *compound-invariant* function[7] of the form

$$w(p) = e^{-\delta(-\ln p)^\gamma}, \tag{9}$$

[6] Henceforth and in the applications, we will refer to this probability distortion as TK function.

[7] In the same article, Prelec derives two other probability weighting functions: the *conditionally-invariant exponential-power* and the *projection-invariant hyperbolic-logarithm* function.

$w(0) = 0$ and $w(1) = 1$. The parameter $\delta \in (0,\ 1)$ governs elevation of the weighting function relative to the 45° line, while $\gamma > 0$ governs curvature and the degree of sensitivity to extreme results relative to medium probability outcomes. When $\gamma < 1$, one obtains an inverse-S shape function. In this model, the parameter δ influences the tendency of over- or underweighting the probabilities, but it has no direct meaning.

With $\delta = 1$, one obtains the more parsimonious single parameter version of Prelec's function:

$$w(p) = e^{-(-\ln p)^{\gamma}}. \tag{10}$$

Note that, in this case, the unique solution of equation $w(p) = p$ for $p \in (0, 1)$ is $p = 1/e$ and elevation of the function does not depend on γ, which governs curvature only.

Function (10) is applied in numerous studies, often as an alternative to function (8). Figure 1 shows some examples of these two weighting functions; both of them depend on a single parameter γ: for lower values of γ the functions exhibit higher curvature.

In the applications, we use the parameters estimated by Tversky and Kahneman [20] for function (8): $\gamma^{+} = 0.61$ and $\gamma^{-} = 0.69$, for w^{+} and w^{-}, which denote the weighting function for probabilities of gains and losses, respectively. The same set of parameters is used also for the Prelec's function. For direct comparison, Fig. 2 shows the TK weighting function (8) and the Prelec's weighting function (10) for $\gamma = 0.61$. It is worth noting that, for this choice of the parameter γ, the TK function displays higher curvature and lower elevation; moreover, extreme sensitivity is slightly higher for the Prelec function.

A parametric function of particular interest is the *switch-power weighting function* proposed by Diecidue et al. [10], which consists in a power function for probabilities below a certain value $\hat{p} \in (0, 1)$ and a dual power function for probabilities above $\hat{p}$; formally w is defined as follows:

$$w(p) = \begin{cases} cp^{a} & \text{if } 0 \le p \le \hat{p}, \\ 1 - d(1-p)^{b} & \text{if } \hat{p} < p \le 1, \end{cases} \tag{11}$$

with five parameters a, b, c, d, and $\hat{p}$. All the parameters are strictly positive, assuming continuity and monotonicity of w. When $\hat{p}$ approaches 1 or 0, w reduces to a power or a dual power probability weighting function, respectively. The authors provide preference foundation for such a family of parametric weighting functions and inverse-S shape under Rank-Dependent Utility based on testable preference conditions.

By assuming differentiability, the number of parameters in (11) reduces to three. For $a, b \le 1$, the function w is concave on $(0, \hat{p})$ and convex on $(\hat{p}, 1)$ (it has an inverse-S shape), while for $a, b \ge 1$ the weighting function is convex for $p < \hat{p}$ and concave for $p > \hat{p}$ (it has an S-shape). In this case, both parameters a and b govern the curvature of w when $a \neq b$. In particular, parameter a describes probabilistic risk attitude for small probabilities; whereas parameter b describes probabilistic risk attitude for medium and large probabilities.

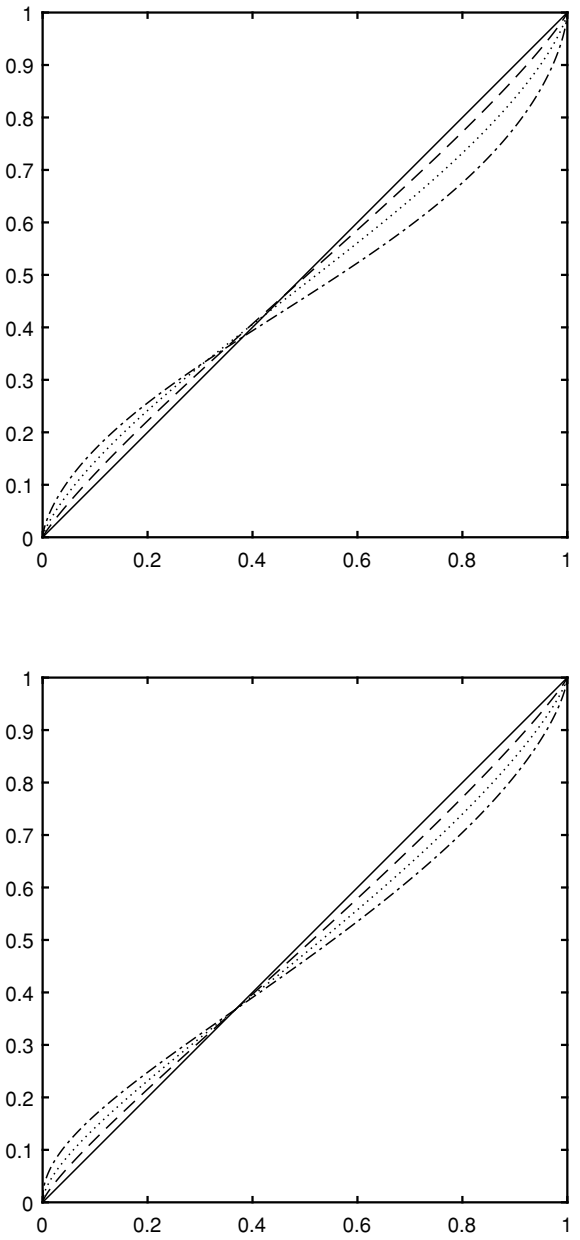


Fig. 1 Examples of the probability weighting functions used by Tversky and Kahneman (upper panel) and Prelec (lower panel); with $\gamma = 0.7, 0.8, 0.9$ (the solid line represents objective probability). Lower values of γ imply higher curvature

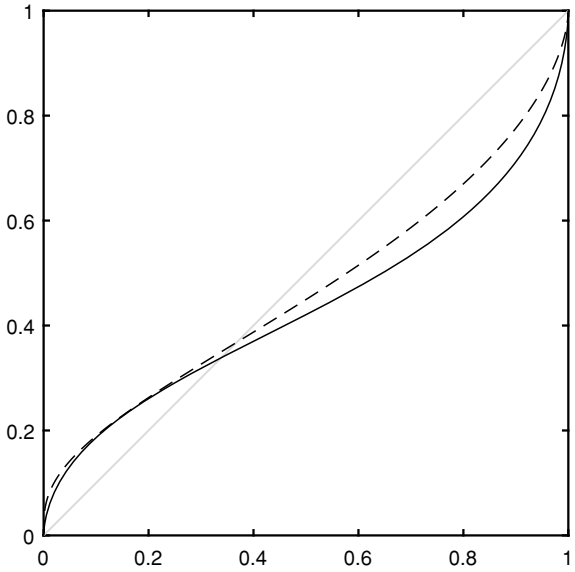


Fig. 2 A comparison between the TK (8) (solid line) and the Prelec (10) (dashed line) probability weighting functions, with $\gamma = 0.61$. The grey line represents objective probability

When $a \neq b$, parameter $\hat{p}$, which indicates the point where probabilistic risk attitudes change from risk aversion to risk seeking (for an inverse-S shape function), may not lie on the 45° line, hence it has not the meaning of dividing the region of over- and underweighting of the probability.

When $a = b$, one obtains a two parameter probability weighting function, which intersects the 45° line at $\hat{p}$. The parameter $\hat{p}$ separates the regions of over- and underweighting of probabilities. If we denote $\delta = \hat{p}$ and $a = \gamma$, the result is the *Constant Relative Sensitivity* (CRS) weighting function considered by Abdellaoui et al. [3].

$$w(p) = \begin{cases} \delta^{1-\gamma} p^{\gamma} & \text{if } 0 \leq p \leq \delta, \\ 1 - (1-\delta)^{1-\gamma}(1-p)^{\gamma} & \text{if } \delta < p \leq 1, \end{cases} \tag{12}$$

with $\gamma > 0$ and $\delta \in [0, 1]$. For $\gamma < 1$ and $0 < \delta < 1$, it has an inverse-S shape. The derivative of w at δ equals γ; this parameter controls for the curvature of the weighting function. The parameter δ indicates whether the interval for overweighting probabilities is larger than the interval for underweighting, and therefore controls for the elevation. Hence, this family of weighting functions allows for a separate and direct modeling of these two features.

Figure 3 shows the plots of the CRS weighting function for different values of the parameters. In the applications, we will adopt also this function, using the TK parameters for γ and, in order to direct compare the CRS function with the single parameter Prelec's function, we set $\delta = 1/e$.

Remember that a convex (concave) weighting function characterizes probabilistic risk aversion (risk proneness), whereas a linear weighting function characterizes probabilistic risk neutrality. Then, the role of δ is to demarcate the interval of prob-

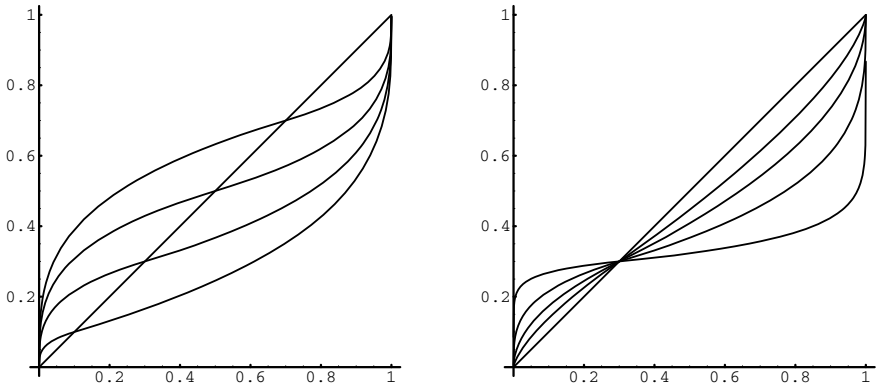


Fig. 3 Two parameter CRS probability weighting function with different elevation (left) for $\gamma = 0.3$ and $\delta = 0.1, 0.3, 0.5, 0.7$ (for higher values of δ the function is more elevated), and curvature (right) for $\delta = 0.3$ and $\gamma = 0.1, 0.3, 0.5, 0.7$ (for lower values of γ the function exhibits higher curvature)

ability risk seeking from the interval of probability risk aversion. In such a case, overweighting corresponds to risk seeking (or optimism) and underweighting corresponds to risk proneness (or pessimism), whereas elevation represents the relative strength of optimism versus pessimism, hence it is a measure of relative optimism, and δ may be interpreted as an index of relative optimism.

Abdellaoui [3] and Gonzalez and Wu [12] find that the weighting function tends to be more elevated for losses than for gains, and that the relative index of optimism for gains (δ^+) is lower than the relative index of pessimism for losses (δ^-).

Curvature is a measure of the degree of sensitivity to changes from impossibility to possibility, it represents the diminishing effect of optimism and pessimism when moving away from extreme probabilities 0 and 1. Hence, parameter γ, controlling for curvature, measures the relative sensitivity of the weighting function. This suggests an interpretation for such a parameter as a measure of relative risk aversion. The *index of relative sensitivity* for a probability weighting function is defined as $-p\frac{\partial^2 w(p)}{\partial p^2} \Big/ \frac{\partial w(p)}{\partial p}$, for $p \in (0, \delta]$, and $-(1-p)\frac{\partial^2 (1-w(p))}{\partial (1-p)^2} \Big/ \frac{\partial (1-w(p))}{\partial (1-p)}$, for $p \in (\delta, 1)$. For function (12), such an index is constant and equals $1 - \gamma$, hence the name of the CRS function.

The possibility of modelling separately curvature and elevation is of particular interest, with potential applications in finance[8] and in particular in portfolio selection, as it allows for direct modeling probabilistic pessimism and optimism of the investor.

Remember that in CPT individuals apply decision weights to ranked outcomes, and in (2) the weights π_i are computed through function w not on probabilities of a single outcome, but they are differences in transformed cumulative (countercumulative) probabilities of losses (gains). In order to understand the dependence of the decision weights π on rank r, and the impact of the shape of the weighting

[8] The CRS weighting function has been adopted by Tversky and Kahneman [16] in a behavioral model for the evaluation of European options.

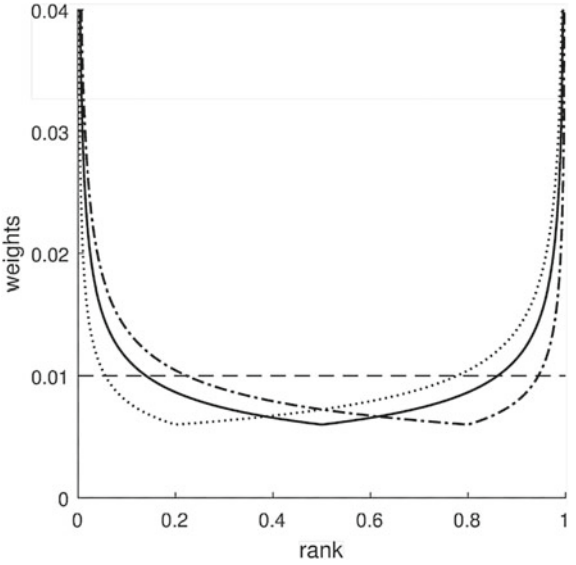


Fig. 4 Dependence of the decision weights on rank r (from best to worst) for w defined as in Eq. (12), with weights $w(p+r) - w(r) \approx p\, w'(r)$ for objective probability $p = 0.01$ (the dashed line depicts neutral psychology). The parameters of probability weighting w are: $\gamma = 0.6$, letting elevation vary, $\delta = 0.2$ (the dotted line curve illustrates relative pessimism), $\delta = 0.5$ (the solid line illustrates that good and bad outcomes are overweighted while intermediate results are underweighted) and $\delta = 0.8$ (the dashed-dotted curve illustrates relative optimism)

function, let us observe Fig. 4. Once the outcomes are ordered (e.g. from best to worst) and cumulative probabilities r are assigned to each ranked result, the decision weights $w(p+r) - w(r)$ can be approximated by $p\, w'(r)$ for objective individual probability p (such a probability is represented by the dashed line which depicts neutral psychology).

In Fig. 4, the weighting function defined in Eq. (12) is used, and results are ordered from best to worst. The curvature parameter is $\gamma = 0.6$, letting vary the elevation parameter. For $\delta = 0.2$, the dotted curve illustrates relative pessimism (probabilities of extreme good outcomes and small probabilities of bad outcomes are overweighted), for $\delta = 0.5$ the solid line illustrates that good and bad outcomes are overweighted, while intermediate results are underweighted, and for $\delta = 0.8$ the dashed-dotted curve illustrates relative optimism (small probabilities of good outcomes are overweighted, probabilities of extreme bad outcomes are overweighted). As γ approaches the value 1 (for lower curvature), the weights tend to the objective probabilities.

Hence, a "small" probability $p = 0.01$ is perceived differently by the PI, depending on the rank of the outcomes. Low probabilities of extreme outcomes are overweighted with respect to a 1% probability of a medium result, but with a different effect for "good" or "bad" extreme outcomes, depending on the probabilistic optimism or pessimism of the individual.

4 A Behavioral CPT Portfolio Selection

Similarly to what done in [5], we assume that a PI selects the portfolio weights in order to maximize her prospect value subject to the usual budget constraint and short selling restrictions. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the vector of portfolio weights, such that $x_j \geq 0$ $(j = 1, 2, \ldots, n)$ and $\sum_{j=1}^{n} x_j = 1$. Let us consider m possible scenarios, with r_{ij} the return of equity j in scenario i, and p_i the probability of each i. In this work we considered equally probable scenarios, i.e. $p_i = 1/m$ for all i.

The portfolio returns, measured relative to a fixed reference point r_0, are the results subjectively evaluated by the PI, with decision weights computed through one of the probability weighing functions discussed in the previous section.[9] Formally, the BP selection model is defined as:

$$\begin{aligned} \max_{\mathbf{x}} \; & \sum_{i=1}^{m} \pi_i \cdot v\left(\sum_{j=1}^{n} x_j r_{ij} - r_0\right) \\ \text{s.t.} \; & \sum_{j=1}^{n} x_j = 1 \\ & x_j \geq 0, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{13}$$

This optimization problem is highly non-linear and non-differentiable so it cannot be easy solved applying traditional optimization techniques. For these reasons, according to what already done in [5], we adopt a solution approach based on the metaheuristic PSO, discussed in the next section.

5 A Particle Swarm Optimization Solution Approach

PSO is an iterative bio-inspired population-based metaheuristic for the solution of global unconstrained optimization problems. In this section, first we introduce the basics of standard PSO, then we present the implementation performed in order to take into account the presence of constraints, as optimization problem (13) is global constrained.

The basic idea of PSO is to replicate the social behavior of shoals of fish or flocks of birds cooperating in the pursuit of a given goal. To this purpose, each member—a *particle*—of the shoal/flock—the *swarm*—explores the search area keeping memory of its best position reached so far, and it exchanges this information with the neighbours in the swarm. Thus, the whole swarm tends to converge towards the best global position reached by the particles.

Such an idea may be formulated as follows: for an optimization problem, every particle of the swarm represents a possible solution. Initially, each particle is assigned

[9] In particular, in the applications we adopt the TK, Prelec, and CRS functions.

to a random position, $\mathbf{x}_j^1$, and to a random velocity, $\mathbf{v}_j^1$. In a few details, let us consider the global optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is the objective function. Suppose we apply PSO for its solution, where M particles are considered. At the k-th iteration of the algorithm, four vectors are associated to the j-th particle, with $j = 1, \ldots, M$: $\mathbf{x}_j^k \in \mathbb{R}^d$, which is its current position; $\mathbf{v}_j^k \in \mathbb{R}^d$, which is its current velocity; $\mathbf{p}_j \in \mathbb{R}^d$, which is its best position visited so far; $\mathbf{p}_g$, which is the best position visited so far by the swarm. Moreover, $pbest_j = f(\mathbf{p}_j)$ is the value of the objective function in $\mathbf{p}_j$, and $gbest = f(\mathbf{p}_g)$ is the value of the objective function in $\mathbf{p}_g$.

The algorithm for a minimization problem, in the version with *inertia weights* h, which is the one we use, is reported in the following:

1. Set $k = 1$ and evaluate $f(\mathbf{x}_j^k)$, for $j = 1, \ldots, M$. Set $pbest_j = +\infty$ for $j = 1, \ldots, M$, and $gbest = +\infty$.
2. If $f(\mathbf{x}_j^k) < pbest_j$, then set $\mathbf{p}_j = \mathbf{x}_j^k$ and $pbest_j = f(\mathbf{x}_j^k)$. If $f(\mathbf{x}_j^k) < gbest$, then set $\mathbf{p}_g = \mathbf{x}_j^k$ and $gbest = f(\mathbf{x}_j^k)$.
3. Update position and velocity of the j-th particle, with $j = 1, \ldots, M$, as

$$\begin{cases} \mathbf{v}_j^{k+1} = h\mathbf{v}_j^k + c_1\mathbf{U}_{[0,1]}(\mathbf{p}_j - \mathbf{x}_j^k) + c_2\mathbf{U}_{[0,1]}(\mathbf{p}_g - \mathbf{x}_j^k) \\ \mathbf{x}_j^{k+1} = \mathbf{x}_j^k + \mathbf{v}_j^{k+1} \end{cases},$$

where h, c_1 and c_2 are the inertia weight, the cognitive constant and the social constant, respectively, and $\mathbf{U}_{[0,1]} \in \mathbb{R}^d$ and its components are uniformly randomly distributed in [0, 1].
4. If a pre-established convergence criterion is not satisfied, then set $k = k + 1$ and go to step 2.

Note that the values of h, c_1 and c_2 affect the strength of the attractive forces towards $\mathbf{v}_j^k$, $\mathbf{p}_j$ and $\mathbf{p}_g$, respectively. In order to get the convergence of the swarm, c_1 and c_2 have to be set carefully in accordance with the value of h.

To deal with the presence of constraints, different strategies are proposed in the literature to ensure that feasible positions are generated at any iterations of PSO. However, in this paper we use PSO accordingly to the original intent, that is as a tool for the solution of unconstrained optimization problems. To this purpose, we reformulate problem (13) into an unconstrained one using a nondifferentiable penalty function method already applied in the financial context [8, 9]. Such an approach is known as *exact penalty method*, where the term "exact" refers to the correspondence between the optimizers of the original constrained problem and the optimizers of the unconstrained (penalized) one.

The reformulated version of BP optimization problem (13) is then

$$\max_{\mathbf{x}} \sum_{i=1}^{m} \pi_i \cdot v\left(\sum_{j=1}^{n} x_j r_{ij} - r_0\right) - \frac{1}{\epsilon}\left[\left|\sum_{j=1}^{n} x_j - 1\right| + \sum_{j=1}^{n} \max\left(0, -x_j\right)\right], \quad (14)$$

Table 1 Statistics and Sharpe ratio for the weekly returns of the Index and ot the 10 sectorial indices in the starting in-sample period, January 2001–June 2018

	Mean	Std. dev.	Skewness	Kurtosis	Sharpe r.
Index	0.00008	0.02632	−1.11147	9.89582	0.00308
Oil & gas	0.00001	0.03239	−0.63774	6.94830	0.00038
Basic materials	0.00098	0.03562	−0.63033	5.31644	0.02743
Industrials	0.00059	0.03021	−0.57849	4.35563	0.01950
Consumer goods	0.00118	0.02327	−0.56259	5.72631	0.05049
Health care	0.00053	0.02422	−0.71067	10.27270	0.02176
Consumer services	−0.00004	0.02539	−0.90060	6.82485	−0.00168
Telecommunications	−0.00074	0.02861	−0.69071	4.60984	−0.02573
Utilities	0.00005	0.02595	−1.81352	19.26360	0.00184
Financials	−0.00072	0.03632	−0.94259	8.66486	−0.01983
Technology	−0.00050	0.03870	−0.62434	3.33019	−0.01286

where ϵ is the so-called penalty parameter. Note that a correct setting of ϵ ensures the correspondence between the solutions of the original constrained problem and problem (14).

Finally, note that PSO is a stochastic optimizer due to the random initialization of the positions and velocities of the particles; and due to the presence of uniformly randomly distributed variables in the updating formula of the particles velocities. To manage such stochasticity of PSO, the algorithm is usually let run several times; then, among all the obtained candidate solutions, the one with the lowest fitness is chosen as "optimal".

6 An Application to the European Stock Market

In order to assess the CPT portfolio selection model, especially the effect of different probability distortion functions, we carried out an analysis based on the European equity market. The benchmark used for such market is the STOXX Europe 600 Index; the investible assets are represented by the ten sectorial indices (Oil & gas, Basic materials, Industrials, Consumer goods, Health care, Consumer services, Telecommunications, Utilities, Financials, Technology).

In Table 1, we report statistics and the Sharpe ratio for the weekly returns of the Index, we consider as benchmark, and of the ten sectorial indices in the starting in-sample period, January 2001–June 2018.

We compare, in an out-of-sample analysis, the performances and risk profiles of portfolios obtained applying three alternative probability distortion functions (TK, Prelec, and CRS); for each of them, two different values for the reference point are considered: $r_0 = 0\%$ and $r_0 = 2.5\%$ (annualized values).

Table 2 Statistics for the out-of-sample returns of the optimal BPs. The pedix indicates the annualized value of the reference point

	Index	$TK_{0\%}$	$TK_{2.5\%}$	$Pr_{0\%}$	$Pr_{2.5\%}$	$CRS_{0\%}$	$CRS_{2.5\%}$
Mean	0.00013	0.00127	0.00188	0.00122	0.00098	0.00143	0.00127
Std. dev.	0.01757	0.01648	0.01675	0.01676	0.01677	0.01711	0.01737
Skewness	−0.48461	−0.33057	−0.57295	−0.04329	−0.02054	−0.33092	−0.29151
Kurtosis	3.00339	2.85808	2.78287	2.45395	2.35847	2.72909	2.73855
Sharpe ratio	0.00730	0.07735	0.11210	0.07302	0.05843	0.08378	0.07295

The overall testing period goes from July, 2018 to June, 2019. The out-of-sample analysis is carried out with a rolling window procedure with weekly updating. At each step, a set of 200 equally-probable scenarios is generated applying historical bootstrapping over an in-sample period of past realized returns starting from January 2001 for the benchmark and for the sectorial indices. The obtained scenarios are then used as input in the optimization procedure. The optimal portfolio composition is then evaluated, out-of-sample at the realized market returns of the subsequent week. At each step the in-sample period is updated shifting by one week, dropping the oldest realizations of the returns and adding the new observed ones. The optimization is re-run and a new optimal portfolio composition is obtained and evaluated at realized returns. This scheme is then repeated until the entire 1-year out-of-sample evaluation period is covered.

Overall, for the empirical applications, 306 optimization problems have been considered[10] and solved. As for the quality of the optimal solutions obtained for the problems of type (13), this is generally rather satisfactory. Indeed, for each of these optimization problems: first, the candidate solutions produced in the various runs generally displayed similar fitnesses, meaning that the stochasticity of PSO has been well managed; second, the budget constraint is substantially satisfied, being its largest violation, in absolute value, lower than 10^{-16}, as to say that at most 2 euros have not been invested over a starting capital of 1 million of billion euros; third and last, all the no-short selling constraints are always satisfied.

In Table 2, we show the statistics related to the out-of-sample returns achieved by the optimal BPs; in Fig. 5, we provide the relative frequencies for the same returns.

From Table 2, we can observe that, for all the three different probability weighting functions considered, i.e. TK, Prelec and CRS, and for both reference points, $r_0 = 0\%$ and $r_0 = 2.5\%$, the risk/return profiles of the optimized BP outperform the benchmark. In particular, it is worth noting that, over the considered 1-year out-of-sample period, the BP portfolios exhibit higher mean returns with comparable standard deviation. Overall the returns exhibit a slightly negative skewness, with almost symmetric distributions in the case of Prelec probability weighting function.

[10] Three probability weighting functions, times two reference points, times fiftyone out-of-sample weeks.

Table 3 Ratios for the out-of-sample returns of the optimal BPs and the corresponding measures for the Index

	Mean	Std. dev.	Skewness	Kurtosis	Sharpe r.
$TK_{0.0\%}$	9.93	0.94	0.68	0.95	10.59
$TK_{2.5\%}$	14.64	0.95	1.18	0.93	15.35
$Pr_{0.0\%}$	9.54	0.95	0.09	0.82	10.00
$Pr_{2.5\%}$	7.64	0.95	0.04	0.79	8.00
$CSR_{0.0\%}$	11.17	0.97	0.68	0.91	11.47
$CSR_{2.5\%}$	9.87	0.99	0.60	0.91	9.99

When kurtosis is considered, in all the cases the returns of BPs show a lower kurtosis when compared with the Index and, in particular, the Prelec cases are the ones that registered the lower values (see also Fig. 5).

In Table 3, the moments and the Sharpe ratios, for each choice of the probability weighting function and reference point, are compared with the corresponding measures of the Index portfolio, computing their ratios.

With reference to the first column, we can observe that the mean of the BPs, in all cases, is considerably higher than the mean of the Index. Furthermore, to this improved performance corresponds an analogous level of risk; as a result, the Sharpe ratios of the optimally chosen BPs are consistently higher than the Sharpe ratio of the Index.

In Fig. 5, the relative frequencies of the out-of-sample returns for the optimal BPs are presented. In the first column, the returns' distributions for the three different probability weighting functions with reference point $r_0 = 0\%$ are compared, whilst the second column displays an analogous comparison for the cases with reference point $r_0 = 2.5\%$.

Finally, Fig. 6 shows the equity lines for the optimal BPs and for the STOXX 600 Europe Index along the 1-year out-of-sample period, using as starting capital $C = 100$. In the upper graph, the cases corresponding to a $r_0 = 0\%$ reference point are plotted, while in the bottom graph the same comparison is carried out for the cases with reference point $r_0 = 2.5\%$. This second reference point results to be rather demanding given the overall market conditions in the considered testing periods.

In general, the up- and down-trends of the market strongly affect the out-of-sample performances of the optimal portfolios, which behave similarly.

As already noticed, in Tables 2 and 3, in all the considered cases we can observe that the BPs outperformed in the out-of-sample considered period the Index. However, the BPs respond differently to various market phases in correspondence of the two reference points. There is no constant preference ordering among the three different cases, TK, Prelec and CRS.

In the case $r_0 = 0.0\%$ (upper panel in Fig. 6), all three portfolios similarly reduce the losses in the down-trend market in the first part of the out-of-sample period, while in the second part of the period, in an up-trend market, the portfolios based

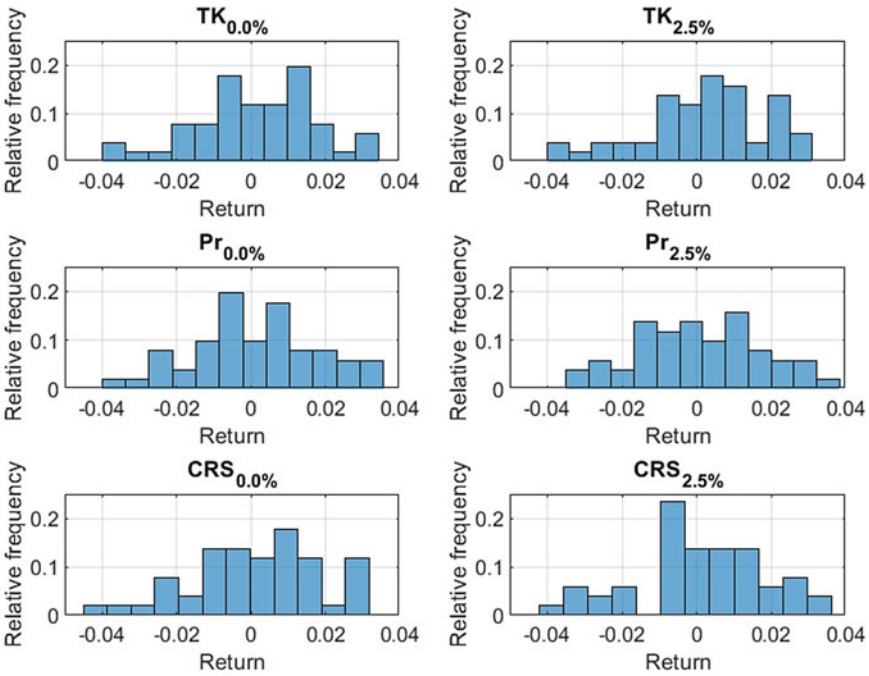


Fig. 5 Relative frequencies of the out-of-sample returns of the optimal BPs

on Prelec and CRS distortion functions seem to perform better (see Fig. 6). For the reference point $r_0 = 2.5\%$ (bottom panel in Fig. 6), we can observe that the volatility of the returns slightly increases for all cases and the portfolios based on the TK distortion function consistently dominate the other portfolios (see also the Sharpe ratio in Tables 2 and 3).

7 Conclusions

In this paper, we proposed a portfolio selection model based on CPT, focusing on the effects on the portfolio risk and return performances of three alternative probability distortion functions: the TK, Prelec and CRS ones, for two different reference points, $r = 0.0\%$ and $r = 2.5\%$, respectively.

The preliminary evidences confirm the effectiveness of the BPs in accommodating for preferences that takes into account loss aversion producing improved risk/return profiles.

In all the considered cases in the tested out-of-sample experiment the optimized BP performed better than the Index both in terms of expected return and Sharpe

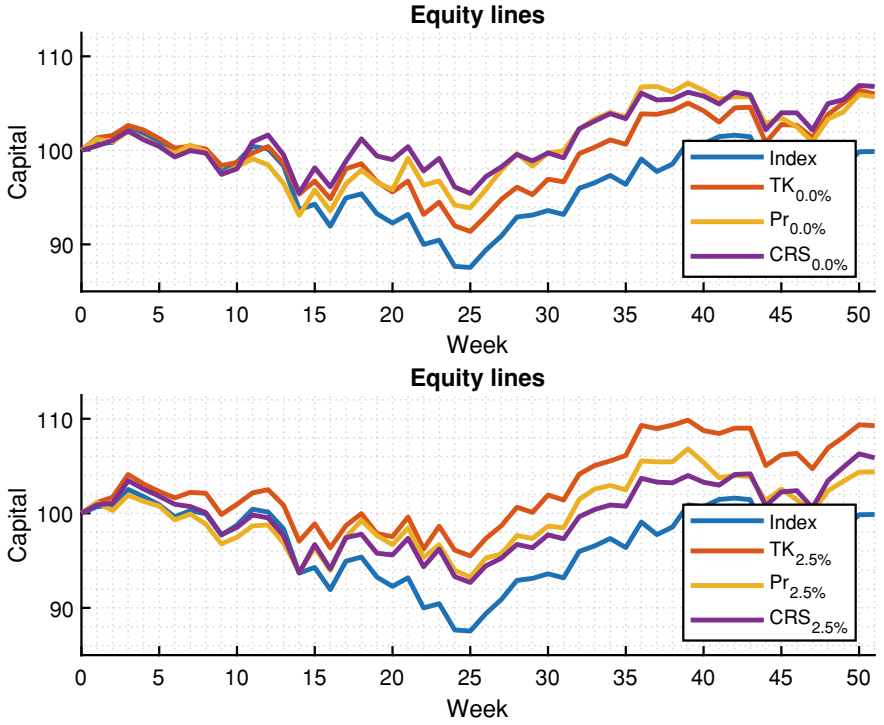


Fig. 6 Out-of-sample equity lines for the Index and the optimal BPs based on different probability weighting functions (TK, Prelec, and CRS functions), with reference point $r_0 = 0\%$ (upper panel) and $r_0 = 2.5\%$ (lower panel)

ratios. As for the composition of the various BPs, we highlight that the selected ones result to be enough diversified.

The role played by the biasing of the objective probabilities introduced by the different probability weighting functions requires further testing to assess the effects of the different components of the distortion (curvature an elevation), also with reference to different market phases and levels of loss aversion.

The choice of the probability weighting function should be driven by the following motivations: its empirical properties, intuitive and empirically testable preference conditions, nonlinear behavior of the probability weighting function. Moreover, a parametric probability weighting function should be parsimonious (remaining consistent with the properties suggested by empirical evidence), in particular when one considers different parameters for the weighting of probability of gains and losses.

Furthermore, future investigations will consider also a sensitivity analysis on the parameters involved both in the value function and in the probability weighting ones.

References

1. Abdellaoui, M.: Parameter-free elicitation of utility and probability weighting functions. Manag. Sci. **46**, 1497–1512 (2000)
2. Abdellaoui, M., Barrios, C., Wakker, P.P.: Reconciling introspective utility with revealed preference: experimental arguments based on prospect theory. J. Econ. **138**, 336–378 (2007)
3. Abdellaoui, M., L'Haridon, O., Zank, H.: Separating curvature and elevation: a parametric probability weighting function. J. Risk Uncertain. **41**, 39–65 (2010)
4. Allais, M.: The general theory of random choices in relation to the invariant cardinal utility function and the specific probability function. The $(U, \theta)-$Model: a general overview. In: Munier, B.R. (ed.) Risk, Decision and Rationality, pp. 231–289. D. Reidel Publishing Company, Dordrecht, Holland (1988)
5. Barro, D., Corazza, M., Nardon, M.: Behavioral aspects in portfolio selection. In: Corazza, M., Gilli, M., Perna, C., Pizzi, C., Sibillo, M. (eds.) Mathematical and Statistical Methods for Actuarial Sciences and Finance (2021)
6. Bleichrodt, H., Pinto, J.L.: A parameter-free elicitation of the probability weighting function in medical decision analysis. Manag. Sci. **46**, 1485–1496 (2000)
7. Bleichrodt, H., Pinto, J.L., Wakker, P.P.: Making descriptive use of prospect theory to improve the prescriptive use of expected utility. Manag. Sci. **47**, 1498–1514 (2001)
8. Corazza, M., di Tollo, G., Fasano, G., Pesenti, R.: A novel hybrid PSO-based metaheuristic for costly portfolio selection problems. Ann. Oper. Res. **304**, 109–137 (2021)
9. Corazza, M., Fasano, G., Gusso, R.: Particle swarm optimization with no-smooth penalty reformulation, for a complex portfolio selection problem. Appl. Math. Comput. **224**, 611–624 (2013)
10. Diecidue, E., Schmidt, U., Zank, H.: Parametric weighting functions. J. Econ. Theory **144**(3), 1102–1118 (2009)
11. Goldstein, W.M., Einhorn, H.J.: Expression theory and the preference reversal phenomena. Psychol. Rev. **94**(2), 236–254 (1987)
12. Gonzalez, R., Wu, G.: On the shape of the probability weighting function. Cognit. Psychol. **38**, 129–166 (1999)
13. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**, 263–291 (1979)
14. Karmarkar, U.S.: Subjectively weighted utility: a descriptive extension of the expected utility model. Organ. Behav. Hum. Perform. **21**, 61–72 (1978)
15. Markowitz, H.: Portfolio selection. J. Fin. **7**, 77–91 (1952)
16. Nardon, M., Pianca, P.: European option pricing under cumulative prospect theory with constant relative sensitivity probability weighting functions. Comput. Manag. Sci. **16**, 249–274 (2018)
17. Prelec, D.: The probability weighting function. Econometrica **66**, 497–527 (1998)
18. Quiggin, J.: A theory of anticipated utility. J. Econ. Behav. Organ. **3**, 323–343 (1982)
19. Shefrin, H., Statman, M.: Behavioral portfolio theory. J. Fin. Quant. Anal. **35**, 127–151 (2000)
20. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of the uncertainty. J. Risk Uncertain. **5**, 297–323 (1992)
21. Wakker, P.P.: Prospect Theory: For Risk and Ambiguity. Cambridge University Press, Cambridge (2010)
22. Wu, G., Gonzalez, R.: Curvature of the probability weighting function. Manag. Sci. **42**(12), 1676–1690 (1996)

Bayesian Quantile Estimation in Deconvolution

Catia Scricciolo

Abstract Estimating quantiles of a population is a fundamental problem of high practical relevance in nonparametric statistics. This chapter addresses the problem of quantile estimation in deconvolution models with known error distributions taking a Bayesian approach. We develop the analysis for error distributions with characteristic functions decaying polynomially fast, the so-called ordinary smooth error distributions that lead to mildly ill-posed inverse problems. Using Fourier inversion techniques, we derive an inequality relating the sup-norm distance between mixture densities to the Kolmogorov distance between the corresponding mixing cumulative distribution functions. Exploiting this smoothing inequality, we show that a careful choice of the prior law acting as an efficient approximation scheme for the sampling density leads to adaptive posterior contraction rates to the regularity level of the latent mixing density, thus yielding a new adaptive quantile estimation procedure.

Keywords Bayesian quantile estimation · Deconvolution · Ordinary smooth error distribution · Mixture model · Posterior distribution · Rate of convergence

1 Introduction

Quantile estimation is a fundamental problem in nonparametric statistics from both the methodological and practical points of view. Estimated quantiles are relevant in applications. However, since quantiles depend nonlinearly on the underlying distribution, it is not always clear how to estimate them, even more in deconvolution problems, see, e.g., § 1.1.2 in [11], pp. 13–14, and the monograph of [14], where observations are affected by additive measurement errors that should be taken into account, otherwise quantile estimates based on the observed measurements would be biased. For example, since high blood pressure can cause cardiovascular diseases, it is important to determine reference values, in particular, percentiles of systolic

C. Scricciolo (✉)
Dipartimento di Scienze Economiche, Università degli Studi di Verona, Polo Universitario Santa Marta,Via Cantarane 24,37129 Verona, VR, Italy
e-mail: catia.scricciolo@univr.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_10

and diastolic blood pressure by features like age, sex etc. The observer is aware that blood pressure is measured with some error due to the lack of precision of the measurement device, which forces to consider indirect observations with implicit measurement errors instead of outcomes of the quantity of interest.

We formally describe the classical convolution model. Let $(X_i)_{i\in\mathbb{N}}$ and $(\varepsilon_i)_{i\in\mathbb{N}}$ be independent sequences of i.i.d. real-valued random variables. Suppose we observe $Y_1, \ldots, Y_n$ such that

$$Y_i = X_i + \varepsilon_i, \quad X_i \perp\!\!\!\perp \varepsilon_i, \quad X_i \overset{\text{iid}}{\sim} \mu_X, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mu_\varepsilon, \quad i = 1, \ldots, n,$$

where Y_i is the signal X_i additively corrupted by the measurement error ε_i, which has density f_ε. If also X_i has density, say f_X, then $f_Y = f_X * f_\varepsilon$. We assume that the error density f_ε is completely known and its Fourier transform $\hat{f}_\varepsilon(t) := \int_{\mathbb{R}} e^{\imath tu} f_\varepsilon(u)\,\mathrm{d}u$, $t \in \mathbb{R}$, verifies the following condition.

Assumption 1.1 The error distribution has finite first moment $\mathbb{E}[|\varepsilon|] < \infty$ and possesses Lebesgue density f_ε with Fourier transform $\hat{f}_\varepsilon$ such that, for constants $\beta,\ R,\ R_1 > 0$,

$$|\hat{f}_\varepsilon(t)|^{-1} \le R(1+|t|)^\beta \quad \text{and} \quad |\hat{f}_\varepsilon^{(1)}(t)| \le R_1(1+|t|)^{-(\beta+1)}, \quad t \in \mathbb{R}. \tag{1}$$

Condition (1), which implies that $\hat{f}_\varepsilon$ decays asymptotically slower than a polynomial, characterizes the so-called ordinary smooth error distributions.

For $\tau \in (0,\ 1)$, let

$$q^\tau = Q(\tau) \equiv F_X^{-1}(\tau) := \inf\{x :\ F_X(x) \ge \tau\}$$

be the τ-quantile of the population X having cumulative distribution function F_X. The problem is to estimate q^τ from indirect observations $Y_1, \ldots, Y_n$. Quantile estimation in deconvolution with measurement error distribution satisfying condition (1) leads to nonlinear functional estimation in a mildly ill-posed inverse problem.

The problem of quantile estimation in deconvolution for the case of known error distribution has been studied by [12], while the more realistic situation in which also the error distribution is unknown and has to be estimated from a sample $\varepsilon_1^*, \ldots, \varepsilon_m^*$ has only recently been investigated by [5]. The former authors proposed a quantile estimator obtained inverting a distribution function estimator constructed using a direct inversion formula instead of integrating the canonical deconvolution density estimator as in [7], which resulted in a non-optimal (in the minimax sense) analysis of the method. Dattner et al. [5], instead, used a plug-in method for distribution function estimation based on a deconvolution density estimator which leads to a minimax-optimal procedure under a local α-Hölder regularity condition on f_X for $\alpha \ge 1/2$, with rates

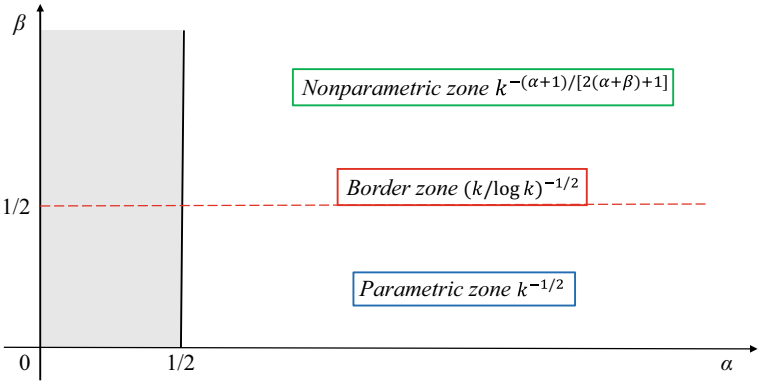


Fig. 1 Partition of the *Regular zone* $\mathscr{Z} = \{(\alpha, \beta) : \alpha \geq 1/2,\ \beta > 0\}$: *Parametric zone* $\{(\alpha, \beta) \in \mathscr{Z} : \beta < 1/2\}$, *Border zone* $\{(\alpha,\ \beta) \in \mathscr{Z} : \beta = 1/2\}$, *Nonparametric zone* $\{(\alpha, \beta) \in \mathscr{Z} : \beta > 1/2\}$

$$\psi_k(\alpha, \beta) := \begin{cases} k^{-1/2}, & \text{for } \beta < 1/2, \\ (k/\log k)^{-1/2}, & \text{for } \beta = 1/2, \\ k^{-(\alpha+1)/[2(\alpha+\beta)+1]}, & \text{for } \beta > 1/2, \end{cases} \qquad \text{where } k := (n \wedge m),$$

which, for $\beta = 1/2$, differ only by a logarithmic factor from the lower bound

$$k^{-(\alpha+1)/[2\alpha+(2\beta\vee 1)+1]}.$$

The existence of different rate *régimes* for $\beta < 1/2$, $\beta = 1/2$ and $\beta > 1/2$ was already pointed out in Theorem 3.2 by [12] and in Theorem 2.1 by [4] for estimating the cumulative distribution function F_X, see Fig. 1. The same distinction also holds when the error distribution is known, with $\psi_n(\alpha, \beta)$ for all $\alpha,\ \beta > 0$.

In this chapter, we consider the inverse problem of estimating single quantiles of the distribution of X taking a Bayesian nonparametric approach. Some results on Bayesian nonparametric quantile estimation in the direct problem, based on a Dirichlet process prior law for the population distribution, were already present in Ferguson's seminal paper [8]. The limiting distribution of the posterior quantile process has been derived by [3], who showed that the posterior law of the rescaled and recentred quantile function converges weakly to a Gaussian process, as the sample size increases. Confidence bands for the quantile function are constructed based on bootstrap approximations of the posterior quantile process. Also the paper by [13] develops and discusses methods for carrying out nonparametric Bayesian inference on the quantile function based on a Dirichlet process prior. The limiting distribution of the quantile process corresponding to a normalized inverse-Gaussian process has been given in [2], see also [1] for the study of the limiting distribution of the quantile process based on prior laws belonging to a general class of popular Bayesian nonparametric priors.

We are aware of no results on Bayesian nonparametric inference for the quantile function in deconvolution problems with known or unknown error distributions. Section 2 contains a general theorem on posterior contraction rates for quantile estimation in deconvolution (Sect. 2.1), together with an inversion inequality (Sect. 2.2). In Sect. 2.3, we apply the general theorem to the case where the noise has Laplace distribution and the prior on the mixing density is a Dirichlet process mixture of Gaussian densities. Auxiliary results are reported in the Appendix.

Notation. In this paragraph, we fix the notation and recall some definitions used throughout the chapter.

- Let $\mathscr{P}$ stand for the set of all probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and $\mathscr{P}_0$ for the subset of Lebesgue absolutely continuous probability measures.
- For a probability measure $P_0 \in \mathscr{P}_0$, let f_0 denote its density. For any $\epsilon > 0$,

$$B_{\mathrm{KL}}(P_0;\, \epsilon^2) := \left\{ P :\, P_0 \left(\log \frac{f_0}{f} \right) \leq \epsilon^2 \right\}$$

  denotes the *Kullback-Leibler* neighborhood of P_0 of radius ϵ^2, where $P_0 g$ stands for the expected value $\int g \, \mathrm{d}P_0$.
- Let $\langle \alpha \rangle$ be the largest integer strictly smaller than $\alpha > 0$. For any interval $I \subseteq \mathbb{R}$ and function g on I, the Hölder norm of g, denoted by $\|g\|_{C^\alpha(I)}$, is defined as

$$\|g\|_{C^\alpha(I)} := \sum_{k=0}^{\langle \alpha \rangle} \|g^{(k)}\|_{L^\infty(I)} + \sup_{x,\, y \in I : x \neq y} \frac{|g^{\langle \alpha \rangle}(x) - g^{\langle \alpha \rangle}(y)|}{|x - y|^{\alpha - \langle \alpha \rangle}}.$$

- Let $C_B(I)$ stand for the set of continuous and bounded functions on I and $C^\alpha(I,\, R) = \{g \in C_B(I) :\, \|g\|_{C^\alpha(I)} \leq R\}$ for the set of continuous and bounded functions on I with Hölder norm uniformly bounded by $R > 0$.
- Let $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$, $x \in \mathbb{R}$, be the density of a standard Gaussian random variable and $\phi_\sigma(\cdot) = \phi(\cdot/\sigma)/\sigma$ its rescaled version.
- For a cumulative distribution function F, we use the notation $b_F := F * K_b - F$ for the *bias* of F, where K_b is a kernel with bandwidth $b > 0$.
- The symbols "$\lesssim$" and "$\gtrsim$" indicate inequalities valid up to a constant multiple that is universal or fixed within the context, but anyway inessential for our purposes.

2 Main Results

This section is devoted to present the main results of the chapter and is split into three parts. The first one contains a general theorem on posterior convergence rates for quantile estimation in deconvolution, the second one reports an instrumental inversion inequality and the third one an application to quantile estimation in deconvolution by a Dirichlet mixture-of-Laplace-normals prior.

Let $Y^{(n)} := (Y_1, \ldots, Y_n)$ be a sample of n i.i.d. observations drawn from the true data generating probability law P_{0Y}, with density $f_{0Y} = f_{0X} * f_\varepsilon$. Given $\tau \in (0, 1)$, let q_{0X}^τ be the τ-quantile of F_{0X}. We want to estimate q_{0X}^τ taking a Bayesian nonparametric approach. Let Π_n be a prior law on a set $\mathscr{P}_1 \subseteq \mathscr{P}_0$ and let $\Pi_n(\cdot \mid Y^{(n)})$ be the resulting posterior

$$\Pi_n(B \mid Y^{(n)}) = \frac{\int_B \prod_{i=1}^n (f_X * f_\varepsilon)(Y_i) \Pi_n(\mathrm{d}\mu_X)}{\int_{\mathscr{P}_1} \prod_{i=1}^n (f_X * f_\varepsilon)(Y_i) \Pi_n(\mathrm{d}\mu_X)}, \quad B \in \mathscr{B}(\mathbb{R}).$$

The goal is to assess the rate $\epsilon_n = o(1)$ for which

$$\Pi_n(|q_X^\tau - q_{0X}^\tau| \leq M\epsilon_n \mid Y^{(n)}) = 1 + o_{\mathbf{P}}(1),$$

where q_X^τ denotes the τ-quantile of the cumulative distribution function F_X associated to f_X and M is a sufficiently large constant.

2.1 Posterior Convergence Rates for Quantiles in Deconvolution

We give general sufficient conditions on the prior law Π_n and the true mixing density f_{0X} so that the posterior leads to an optimal (in the minimax sense) quantile estimation procedure when the error distribution is known and ordinary smooth.

Theorem 1 *Let $\hat{f}_\varepsilon$ verify Assumption 1.1 for $\beta \geq 1$ and $R,\ R_1 > 0$. For $\alpha,\ b,\ r,\ \zeta > 0$, let*

$$\inf_{x \in [-\zeta, \zeta]} f_{0X}(x - q_{0X}^\tau) \geq r \tag{2}$$

and

$$\|b_{F_{0X}}\|_\infty \lesssim b^{\alpha+1}. \tag{3}$$

Suppose that, for a positive sequence $\epsilon_n \to 0$, with $n\epsilon_n^2 \to \infty$, constants $c_1,\ c_2,\ c_3 > 0$ and sets $\mathscr{P}_n \subseteq \mathscr{P}_0$,

$$\begin{aligned} \Pi_n(\mathscr{P} \setminus \mathscr{P}_n) &\leq c_1 \exp(-(c_2 + 4)n\epsilon_n^2), \\ \Pi_n(B_{\mathrm{KL}}(P_{0Y};\ \epsilon_n^2)) &\geq c_3 \exp(-c_2 n\epsilon_n^2) \end{aligned} \tag{4}$$

and there exist constants $C_1,\ \bar{b} > 0$ such that, for every $\mu_X \in \mathscr{P}_n$,

$$\|b_{F_X}\|_1 \leq C_1 b^{\alpha+1} \text{ for all } b \leq \bar{b}. \tag{5}$$

If, for K large enough,

$$\Pi_n(\|f_Y - f_{0Y}\|_\infty > K\epsilon_n \mid Y^{(n)}) = o_{\mathbf{P}}(1), \tag{6}$$

then, for sufficiently large M,

$$\Pi_n(|q_X^\tau - q_{0X}^\tau| > M(\epsilon_n \log n)^{(\alpha+1)/(\alpha+\beta)} \mid Y^{(n)}) = o_{\mathbf{P}}(1).$$

Proof We begin by noting that, while q_{0X}^τ is fixed, the τ-quantile q_X^τ of F_X is a random variable. Since F_X has density f_X, there exists a (random) point q_*^τ between q_{0X}^τ and q_X^τ such that

$$F_X(q_X^\tau) - F_X(q_{0X}^\tau) = f_X(q_*^\tau)[(q_{0X}^\tau \vee q_X^\tau) - (q_{0X}^\tau \wedge q_X^\tau)]. \tag{7}$$

Using (7), we have

$$\begin{aligned} 0 = \tau - \tau = F_X(q_X^\tau) - F_{0X}(q_{0X}^\tau) &= [F_X(q_X^\tau) - F_X(q_{0X}^\tau)] + \underbrace{[F_X(q_{0X}^\tau) - F_{0X}(q_{0X}^\tau)]}_{=:\Delta} \\ &= f_X(q_*^\tau)[(q_{0X}^\tau \vee q_X^\tau) - (q_{0X}^\tau \wedge q_X^\tau)] + \Delta. \end{aligned}$$

Since, as later on shown, with Π_n-probability one, $f_X(q_*^\tau)$ is (uniformly) bounded away from zero, we get that

$$|q_X^\tau - q_{0X}^\tau| = \frac{|\Delta|}{f_X(q_*^\tau)}.$$

Let K_b be a kernel with bandwidth $b > 0$. Then,

$$\begin{aligned} |\Delta| &= |F_X(q_{0X}^\tau) \mp (F_{0X} * K_b)(q_{0X}^\tau) \mp (F_X * K_b)(q_{0X}^\tau) - F_{0X}(q_{0X}^\tau)| \\ &\le |[F_{0X} * K_b - F_{0X}](q_{0X}^\tau)| + |[(F_X - F_{0X}) * K_b](q_{0X}^\tau)| + |[F_X * K_b - F_X](q_{0X}^\tau)| \\ &= |b_{F_{0X}}(q_{0X}^\tau)| + |[(F_X - F_{0X}) * K_b](q_{0X}^\tau)| + |b_{F_X}(q_{0X}^\tau)|. \end{aligned} \tag{8}$$

By condition (3),

$$I := |b_{F_{0X}}(q_{0X}^\tau)| = O(b^{\alpha+1}).$$

By Theorem 2, for $b > 0$ small enough,

$$\begin{aligned} II := |[(F_X - F_{0X}) * K_b](q_{0X}^\tau)| &\le \|(F_X - F_{0X}) * K_b\|_\infty \\ &\lesssim \|F_Y - F_{0Y}\|_\infty + b^{-(\beta-1)_+} \log(1/b) \|f_Y - f_{0Y}\|_\infty. \end{aligned}$$

By condition (5), over $\mathscr{P}_n$,

$$III := |b_{F_X}(q_{0X}^\tau)| = O(b^{\alpha+1}).$$

Combining previous bounds on I, II and III, we have

$$|\Delta| \lesssim b^{\alpha+1} + \|F_Y - F_{0Y}\|_\infty + b^{-(\beta-1)_+} \log(1/b) \|f_Y - f_{0Y}\|_\infty. \tag{9}$$

Conditions in (4) jointly imply that $\Pi_n(\mathscr{P} \setminus \mathscr{P}_n \mid Y^{(n)}) \to 0$ in P_{0Y}^n-probability. Then, by Lemma 1,

$$\Pi_n(\|F_Y - F_{0Y}\|_\infty > M_n \epsilon_n \mid Y^{(n)}) = o_{\mathbf{P}}(1),$$

which, combined with assumption (6) and inequality (9), yields that, for a suitable choice of b,

$$|\Delta| \lesssim (\epsilon_n \log n)^{(\alpha+1)/(\alpha+\beta)}.$$

It remains to be shown that, with Π_n-probability one, $f_X(q_*^\tau)$ is (uniformly) bounded away from zero. For a constant $0 < \eta < r$ (not depending on f_X nor on q_*^τ) and a sufficiently large n so that $\epsilon_n < \eta$, by the convergence in (6), over a set of posterior measure tending to one in P_{0Y}^n-probability, we have

$$\eta > \|f_X - f_{0X}\|_\infty \geq |f_X(x) - f_{0X}(x)| \quad \text{for every } x \in [q_{0X}^\tau - \zeta,\ q_{0X}^\tau + \zeta].$$

Since the interval $[q_{0X}^\tau - \zeta,\ q_{0X}^\tau + \zeta]$ eventually includes both points q_X^τ and q_*^τ, from condition (2) we have

$$f_X(q_*^\tau) > f_{0X}(q_*^\tau) - \eta \geq \inf_{x \in [q_{0X}^\tau - \zeta,\, q_{0X}^\tau + \zeta]} f_{0X}(x) - \eta \geq r - \eta > 0.$$

The assertion follows. □

Remarks

Some remarks and comments on Theorem 1 that aim at spelling out the assumptions used in the proof, as well as at clarifying their role, are in order.

(i) The lower bound on $\hat{f}_\varepsilon$ and the upper bound on $\hat{f}_\varepsilon^{(1)}$ in condition (1) are standard assumptions on the errors' law in deconvolution problems.

(ii) The bias term I in (8) can be controlled using, for instance, a local Hölder smoothness condition on f_{0X} at q_{0X}^τ. If, in fact, for $\alpha,\ \zeta,\ R > 0$, we have

$$f_{0X}(\cdot - q_{0X}^\tau) \in C^\alpha([-\zeta,\ \zeta],\ R),$$

then, by virtue of Lemma 5.2 in [5], condition (3) is verified. Since the quantile function is estimated pointwise, it seems reasonable to describe the smoothness of f_{0X} locally on a Hölder scale and not globally by, for instance, a decay condition on the tails of the Fourier transform of f_{0X}. One such condition could, for example, be the following one: there exists $\alpha > 0$ such that

$$\int_{\mathbb{R}} |t|^{\alpha} |\hat{f}_{0X}(t)| \, \mathrm{d}t < \infty.$$

Then, by Lemma 2, condition (3) is verified.

(iii) Concerning the bias term III, condition (5) is satisfied if, for instance, the (random) density f_X is modelled as a Gaussian mixture, $f_X = \mu_H * \phi_\sigma$. In fact, by Lemma 3, we have $III \leq \|b_{F_X}\|_\infty \leq C_1 b^{\alpha+1}$ for a universal constant $C_1 > 0$ not depending on μ_H.

(iv) General sufficient conditions on the prior law and the data generating probability measure for assessing posterior contraction rates in L^r-metrics, $1 \leq r \leq \infty$, are given in [10], see Theorems 2 and 3, pp. 2891–2892, along with examples of priors attaining minimax-optimal rates. It is important to note that the conditions in (4) verified for a sieve set $\mathscr{P}_n$ of a prescribed form are two of the three main sufficient conditions for deriving sup-norm contraction rates in the above mentioned theorems. A remarkable feature of Theorem 1 is, therefore, the fact that, in order to obtain posterior convergence rates for single quantiles of X, which is an involved mildly ill-posed inverse problem, it is enough to derive posterior contraction rates in sup-norm in the direct problem, which is more gestible. Granted Assumption 1.1, in fact, the essential conditions to verify are those listed in (4), which are (almost) sufficient for the posterior law to contract at rate ϵ_n around f_{0Y}. This simplification is due to an inversion inequality which is presented in Sect. 2.2.

v) If, under Assumption 1.1 on $\hat{f}_\varepsilon$, with $\beta \geq 1$, and some α-regularity condition on f_{0X}, the sup-norm convergence rate ϵ_n in the direct density estimation problem is, up to a logarithmic factor, of the order $O(n^{-(\alpha+\beta)/[2(\alpha+\beta)+1]})$, then quantiles are estimated, up to a log-factor, at the minimax-optimal rate $n^{-(\alpha+1)/[2(\alpha+\beta)+1]}$.

2.2 *Inversion Inequality for the Kolmogorov Metric*

In this section, we present an inversion inequality relating the Kolmogorov distance between the mixing cumulative distribution functions to either the sup-norm distance between the mixture densities or the Kolmogorov distance between the corresponding mixed cumulative distribution functions, when the error density has Fourier transform decaying not faster than polynomially at infinity. This inequality is analogous to that of Theorem 3.2 in [15] and is derived using the same idea, which consists in employing a suitable kernel to smooth the mixing cumulative distribution functions F_X and F_{0X} to then bound the Kolmogorov distance between the smoothed versions, meanwhile controlling the bias induced by the smoothing.

Theorem 2 *Let $\mu_X,\ \mu_{0X} \in \mathscr{P}_0$, with densities $f_X,\ f_{0X}$ such that $\|f_X\|_\infty,\ \|f_{0X}\|_\infty < \infty$. Let $\hat{f}_\varepsilon$ satisfy Assumption 1.1 for $\beta,\ R,\ R_1 > 0$. Then, for the probability measures $\mu_Y := \mu_X * \mu_\varepsilon$, $\mu_{0Y} := \mu_{0X} * \mu_\varepsilon$, with cumulative distribution functions $F_Y,\ F_{0Y}$, respectively, and a sufficiently small $b > 0$, we have*

$$\|F_X - F_{0X}\|_\infty \lesssim b + T,$$

where

$$T \lesssim \|F_Y - F_{0Y}\|_\infty + \begin{cases} b^{-(\beta-1/2)_+} |\log b|^{1+\mathbf{1}_{\{\beta=1/2\}}/2}\, \|F_Y - F_{0Y}\|_\infty \\ \quad or \\ b^{-(\beta-1)_+} |\log b|\, \|f_Y - f_{0Y}\|_\infty. \end{cases} \tag{10}$$

If, in addition, for $\alpha > 0$ the probabilities μ_{0X} and μ_X satisfy conditions (3) and (5), respectively, then

$$\|F_X - F_{0X}\|_\infty \lesssim b^{\alpha+1} + T,$$

with T as in (10).

The proof, which proceeds along the same lines as those of Theorem 3.2 in [15], is omitted to avoid duplications. The inversion inequality holds true under only Assumption 1.1 on $\hat{f}_\varepsilon$ and its derivative $\hat{f}_\varepsilon^{(1)}$, when no smoothness condition is imposed on $f_X,\ f_{0X}$, and jointly with condition (5), when f_{0X} possesses some type of α-regularity so that condition (3) is satisfied.

2.3 Applications

In this section, we present applications of the previous results when the mixing density f_X is modelled as a mixture of Gaussian densities. We consider a Dirichlet process mixture-of-normals as a prior on $f_X = \mu_H * \phi_\sigma$ so that $f_Y = f_X * f_\varepsilon = (\mu_H * \phi_\sigma) * f_\varepsilon$, with μ_H distributed according to a Dirichlet process with base measure H_0, in symbols, $\mu_H \sim \mathscr{D}_{H_0}$, for some finite, positive measure H_0 verifying the following assumption that will be hereafter in force.

Assumption 2.1 The base measure H_0 has a continuous and positive density h_0 on $\mathbb{R}$ such that, for constants $b_0,\ c_0 > 0$ and $0 < \delta \leq 1$,

$$h_0(u) = c_0 \exp\left(-b_0 |u|^\delta\right), \quad u \in \mathbb{R}.$$

As a prior on the bandwidth $\sigma > 0$, we take an inverse-gamma distribution $\mathrm{IG}(\nu,\ \gamma)$, with shape parameter $\nu > 0$ and scale parameter $\gamma = 1$.

The sampling density $f_{0Y} = f_{0X} * f_\varepsilon$ is assumed to be a mixture of Laplace densities, with true mixing density f_{0X} satisfying the following conditions.

Assumption 2.2 There exists $\alpha > 0$ such that

$$\forall\, d = \mp 1/2, \quad \int_{\mathbb{R}} |t|^{2\alpha} |\widehat{e^{d\cdot} f_{0X}}(t)|^2 \, \mathrm{d}t < \infty. \tag{11}$$

Assumption 2.3 There exist $0 < \upsilon \leq 1$, $L_0 \in L^1(\mathbb{R})$ and $R \geq 2m/\upsilon$, for the smallest integer $m \geq [2 \vee (\alpha + 2)/2]$, such that f_{0X} satisfies

$$|f_{0X}(x + \zeta) - f_{0X}(x)| \leq L_0(x)|\zeta|^{\upsilon}, \quad \text{for every } x,\ \zeta \in \mathbb{R}, \tag{12}$$

with envelope function L_0 verifying the following integrability condition

$$\int_{\mathbb{R}} e^{|x|/2} \left(\frac{L_0(x)}{f_{0X}(x)} \right)^R f_{0X}(x) \, \mathrm{d}x < \infty, \tag{13}$$

where, for some constant $C_0 > 0$,

$$f_{0X}(x) \lesssim e^{-(1+C_0)|x|}, \quad x \in \mathbb{R}, \tag{14}$$

and satisfies condition (2).

Condition (11) is a Sobolev-type requirement on the tail decay behaviour of the Fourier transform of f_{0X}, while condition (12) is a local Hölder regularity requirement. Condition (13) is a technical integrability condition paired with the exponential tail decay condition in (14).

Proposition 1 *Let $Y_1, \ldots, Y_n$ be i.i.d. observations from $f_{0Y} := f_{0X} * f_\varepsilon$, where f_ε is the density of a standard Laplace distribution and f_{0X} satisfies Assumptions 2.2 and 2.3. For some constant $\kappa > 0$, let $\epsilon_n = n^{-(\alpha+2)/(2\alpha+5)} (\log n)^\kappa$. Suppose that there exists $\mathscr{P}_n \subseteq \mathscr{P}_0$ such that, for some constant $C > 0$, the prior probability $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \lesssim e^{-Cn\epsilon_n^2}$ and, for K large enough,*

$$\Pi(\mu_X \in \mathscr{P}_n : \|f_X * f_\varepsilon - f_{0Y}\|_\infty > K\epsilon_n \mid Y^{(n)}) = o_{\mathbf{P}}(1). \tag{15}$$

Then, for positive constants M large enough and κ',

$$\Pi(|q_X^\tau - q_{0X}^\tau| > M n^{-(\alpha+1)/(2\alpha+5)} (\log n)^{\kappa'} \mid Y^{(n)}) = o_{\mathbf{P}}(1).$$

Proof We apply Theorem 1. By Assumption 2.2 and Lemma 2, condition (3) is verified. Given the structure of the prior, also condition (5) is satisfied by virtue of Lemma 3. It follows from [15] that the small-ball prior probability estimate in (4) is satisfied for $n^{-(\alpha+2)/(2\alpha+5)}$, up to a log-factor. The convergence in (6) then follows from the hypothesis of prior mass negligibility of the set $\mathscr{P} \setminus \mathscr{P}_n$, combined with assumption (15). The assertion holds. □

While the previous result is based on an application of the sup-norm version of the smoothing inequality in Theorem 2, the result of the following Proposition 2 is an application of the Kolmogorov distance version. We assume that the true mixing density admits a representation as a mixture of normal densities $f_{0X} = \mu_{H_0} * \phi_{\sigma_0}$, with $\sigma_0 > 0$ fixed and μ_{H_0} satisfying the following assumption.

Assumption 2.4 For some constants $c_0 > 0$ and $\varpi \geq 2$, the probability measure μ_{H_0} verifies the tail condition

$$\mu_{H_0}(z : |z| > t) \lesssim \exp(-c_0 t^{\varpi}) \quad \text{as } t \to \infty.$$

Proposition 2 *Let $Y_1, \ldots, Y_n$ be i.i.d. observations from a density $f_{0Y} := f_{0X} * f_\varepsilon$, where f_ε satisfies Assumption 1.1 for some $\beta > 0$ and $f_{0X} = \mu_{H_0} * \phi_{\sigma_0}$, with μ_{H_0} satisfying Assumption 2.4. Then there exist positive constants M' large enough and κ' so that*

$$\Pi(|q_X^\tau - q_{0X}^\tau| > M' n^{-(\alpha+1)/[2\alpha+(2\beta\vee 1)+1]} (\log n)^{\kappa'} \mid Y^{(n)}) = o_{\mathbf{P}}(1).$$

Proof Conditions (2), (3) and (5) are easily checked to be satisfied. The small-ball prior probability estimate in (4) can be shown to be satisfied for ϵ_n equal to $n^{-1/2}$, up to a log-factor, as in Theorem 1 of [16], pp. 486–487 and 513–516. Then, Lemma 1 implies that, for a suitable constant $\kappa'' > 0$, the posterior probability $\Pi(\|F_Y - F_{0Y}\|_\infty > M_n n^{-1/2} (\log n)^{\kappa''} \mid Y^{(n)}) = o_{\mathbf{P}}(1)$. The assertion follows applying the Kolmogorov distance version of the smoothing inequality in Theorem 2. □

The interesting feature of Proposition 2 is that, due to the representation of the true mixing density as a mixture of normals, the prior concentration rate turns out to be nearly parametric, which implies that the posterior contracts at nearly $\sqrt{n}$-rate on Kolmogorov neighborhoods of F_{0Y}, thus allowing, by means of the inversion inequality of Theorem 2, to recover the whole range of rate *régimes* for different values of $\beta > 0$.

3 Final Remarks

In this chapter, we have studied the problem of quantile estimation in deconvolution models with ordinary smooth error distributions, taking a Bayesian nonparametric approach. We have given sufficient conditions on the prior law and the true data generating density so that single quantiles can be estimated at minimax-optimal rates, up to a log-factor. The crucial step is an inversion inequality relating the Kolmogorov distance between the mixing cumulative distribution functions to either the sup-norm distance between the mixture densities or the Kolmogorov distance between the corresponding mixed cumulative distribution functions. Validity of the sup-norm

version is limited to $\beta \geq 1$, while the Kolmogorov distance version covers the whole range of values of $\beta > 0$ and leads to minimax-optimal posterior convergence rates for quantiles, provided that the posterior contracts at a nearly $\sqrt{n}$-rate on Kolmogorov neighborhoods of F_{0Y}, a result that is elusive to us at the moment in all generality. In fact, we could prove it only for the case when the true mixing density admits a representation as a location mixture of Gaussians so that the sampling density is itself a location mixture of normals. Proving the result in general is an open challenging problem.

As a last remark, we note that the derivation of the upper bound on the estimation error in (9) is based on Fourier inversion techniques that are useful to extend results from single quantiles to quantile function estimation in L^1-norm, as shown in [15]. In fact, the L^1-norm between quantile functions coincides with the L^1-norm between the cumulative distribution functions, namely, with the L^1-Wasserstein distance between the associated probability laws.

Acknowledgements The author would like to thank the Editors and two anonymous referees for valuable comments and remarks. She is a member of the *Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni* (GNAMPA) of the *Istituto Nazionale di Alta Matematica* (INdAM).

Appendix

The following lemma provides sufficient conditions on the true cumulative distribution function F_{0Y} and the prior law Π_n so that the posterior measure concentrates on Kolmogorov neighborhoods of F_{0Y}. It is a modification of Lemma 1 in [17], pp. 123–125, and of Lemma B.1 in [15], pp. 24–25, with a weaker condition on the prior concentration rate. In fact, Kullback-Leibler type neighborhoods, which involve also the second moment of the log-ratio $\log(f_{0Y}/f_Y)$, can be replaced by Kullback-Leibler neighborhoods.

Lemma 1 *Let F_{0Y} be a continuous cumulative distribution function. Let Π_n be a prior law on a set $\mathscr{P}_1 \subseteq \mathscr{P}_0$ of probability measures with continuous cumulative distribution functions. If, for a constant $C > 0$ and a positive sequence $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$, we have*

$$\Pi_n(B_{\mathrm{KL}}(P_{0Y};\, \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2),$$

then, for a sequence $M_n := \xi(1-\theta)^{-1}(C/2 + L_n)^{1/2}$, with $\theta \in (0,\, 1)$, $\xi > 1$ and $L_n \to \infty$ such that $L_n^{1/2}\epsilon_n \to 0$, we have

$$\Pi_n(\|F_Y - F_{0Y}\|_\infty > M_n\epsilon_n \mid Y^{(n)}) = o_{\mathbf{P}}(1). \tag{16}$$

Proof By Lemma 6.26 of [9], p. 145, with P_{0Y}^n-probability at least equal to $(1 - L_n^{-1})$, we have

$$\int_{\mathscr{P}_1} \prod_{i=1}^{n} \frac{f_Y}{f_{0Y}}(Y_i)\, \Pi_n(\mathrm{d}\mu_Y) \gtrsim \exp\left(-(C + 2L_n)n\epsilon_n^2\right). \tag{17}$$

Following the proofs of Lemma 1 in [17], pp. 123–125, Lemma B.1 in [15], pp. 24–25, and applying the lower bound in (17), the convergence statement in (16) holds true. □

Remark 1 Lemma 1 shows that, by taking L_n to be a slowly varying sequence, Kullback-Leibler type neighborhoods can be replaced by Kullback-Leibler neighborhoods at the cost of an additional factor in the rate not affecting the power of n, which is thus of the order $L_n^{1/2}\epsilon_n$.

The next lemma assesses the order of the sup-norm of the bias of a cumulative distribution function with density in a Sobolev-type space. It is the sup-norm version of Lemma C.2 in [15], which, instead, considers the L^1-norm.

Lemma 2 *Let F_{0X} be the cumulative distribution function of a probability measure $\mu_{0X} \in \mathscr{P}_0$ with density f_{0X}. Suppose that there exists $\alpha > 0$ such that $\int_{\mathbb{R}} |t|^\alpha |\hat{f}_{0X}(t)|\, \mathrm{d}t < \infty$. Let $K \in L^1(\mathbb{R})$ be symmetric, with $\hat{K} \in L^1(\mathbb{R})$ such that $\hat{K} \equiv 1$ on $[-1,\ 1]$. Then, for every $b > 0$,*

$$\|F_{0X} * K_b - F_{0X}\|_\infty = O(b^{\alpha+1}).$$

Proof Let $b_{F_{0X}} := (F_{0X} * K_b - F_{0X})$ be the bias of F_{0X}. By the same arguments used for the function $G_{2,b}$ in [6], pp. 251–252, we have

$$\begin{aligned}\|b_{F_{0X}}\|_\infty := \sup_{x\in\mathbb{R}} |b_{F_{0X}}(x)| &= \sup_{x\in\mathbb{R}} \left| \frac{1}{2\pi} \int_{|t|>1/b} \exp(-\imath t x) \frac{[1 - \hat{K}(bt)]}{(-\imath t)} \hat{f}_{0X}(t)\, \mathrm{d}t \right| \\ &\leq \frac{1}{2\pi} \int_{|t|>1/b} \frac{|1 - \hat{K}(bt)|}{|(-\imath t)|} |\hat{f}_{0X}(t)|\, \mathrm{d}t,\end{aligned}$$

where the mapping $t \mapsto [1 - \hat{K}(bt)][\hat{f}_{0X}(t)\mathbf{1}_{[-1,\,1]^c}(bt)/t]$ is in $L^1(\mathbb{R})$ by the assumption that $(|\cdot|^\alpha \hat{f}_{0X}) \in L^1(\mathbb{R})$. Note that

$$\begin{aligned}\|b_{F_{0X}}\|_\infty &\leq \frac{1}{2\pi} \int_{|t|>1/b} \frac{|1 - \hat{K}(bt)|}{|(-\imath t)|^{\alpha+1}} \underbrace{|(-\imath t)^\alpha \hat{f}_{0X}(t)|}_{=|\widehat{D^\alpha f_{0X}}(t)|}\, \mathrm{d}t \\ &< \frac{b^{\alpha+1}}{2\pi} \int_{|t|>1/b} [1 + |\hat{K}(bt)|]|t|^\alpha |\hat{f}_{0X}(t)|\, \mathrm{d}t \lesssim b^{\alpha+1} \int_{|t|>1/b} |t|^\alpha |\hat{f}_{0X}(t)|\, \mathrm{d}t \lesssim b^{\alpha+1}\end{aligned}$$

because $\|\hat{K}\|_\infty \leq \|K\|_1 < \infty$. The assertion follows. □

The following lemma establishes the order of the sup-norm of the bias of the cumulative distribution function of a Gaussian mixture, when the mixing distribution is any probability measure on the real line and the scale parameter is chosen as a multiple of the kernel bandwidth, up to a logarithmic factor. It is analogous to Lemma G.1 in [15], p. 46, which, instead, considers the L^1-norm. Both results rely upon the fact that a Gaussian density has exponentially decaying tails.

Lemma 3 *Let F_X be the cumulative distribution function of $\mu_X = \mu_H * \phi_\sigma$, with $\mu_H \in \mathscr{P}$ and $\sigma > 0$. Let $K \in L^1(\mathbb{R})$ be symmetric, with $\hat{K} \in L^1(\mathbb{R})$ such that $\hat{K} \equiv 1$ on $[-1, 1]$. Given $\alpha > 0$ and a sufficiently small $b > 0$, for $\sigma = O(2b|\log b^{\alpha+1}|^{1/2})$, we have*

$$\|F_X * K_b - F_X\|_\infty = O(b^{\alpha+1}).$$

Proof Let $b_{F_X} := (F_X * K_b - F_X)$ be the bias of F_X. Defined for every $b, \sigma > 0$ the function

$$\widehat{f_{b,\sigma}}(t) := \frac{1 - \hat{K}(bt)}{t}\hat{\phi}(\sigma t/\sqrt{2})\mathbf{1}_{[-1,1]^c}(bt), \quad t \in \mathbb{R},$$

since $t \mapsto [\hat{\mu}_H(t)\hat{\phi}(\sigma t/\sqrt{2})]\widehat{f_{b,\sigma}}(t)$ is in $L^1(\mathbb{R})$, arguing as for $G_{2,b}$ in [6], pp. 251–252, we have that

$$\begin{aligned}
\|b_{F_X}\|_\infty := \sup_{x\in\mathbb{R}} |b_{F_X}(x)| &= \sup_{x\in\mathbb{R}} \left| \frac{1}{2\pi} \int_{|t|>1/b} \exp(-\imath t x) \frac{[1 - \hat{K}(bt)]}{(-\imath t)} \hat{\mu}_H(t)\hat{\phi}(\sigma t)\, \mathrm{d}t \right| \\
&= \sup_{x\in\mathbb{R}} \left| \frac{1}{2\pi} \int_{|t|>1/b} \exp(-\imath t x) \hat{\mu}_H(t)\hat{\phi}(\sigma t/\sqrt{2})\widehat{f_{b,\sigma}}(t)\, \mathrm{d}t \right| \\
&= \|\mu_H * \phi_{\sigma/\sqrt{2}} * f_{b,\sigma}\|_\infty,
\end{aligned}$$

where $f_{b,\sigma}(\cdot) := (2\pi)^{-1} \int_{\mathbb{R}} \exp(-\imath t \cdot)\widehat{f_{b,\sigma}}(t)\, \mathrm{d}t$ because $\widehat{f_{b,\sigma}} \in L^1(\mathbb{R})$. Since $\|\mu_H * \phi_{\sigma/\sqrt{2}}\|_1 = 1$ and $\|f_{b,\sigma}\|_\infty \leq \|\widehat{f_{b,\sigma}}\|_1 < \infty$ for all $\mu_H \in \mathscr{P}$ and $\sigma > 0$, by Young's convolution inequality,

$$\|b_{F_X}\|_\infty = \|\mu_H * \phi_{\sigma/\sqrt{2}} * f_{b,\sigma}\|_\infty \leq \|\mu_H * \phi_{\sigma/\sqrt{2}}\|_1 \times \|f_{b,\sigma}\|_\infty = \|f_{b,\sigma}\|_\infty,$$

where

$$\|f_{b,\sigma}\|_\infty \le \|\widehat{f_{b,\sigma}}\|_1 \le \int\limits_{|t|>1/b} \frac{1+|\hat{K}(bt)|}{|t|}\hat{\phi}(\sigma t/\sqrt{2})\,\mathrm{d}t$$
$$\lesssim b \int\limits_{|t|>1/b} \hat{\phi}(\sigma t/\sqrt{2})\,\mathrm{d}t \lesssim (b/\sigma)^2\hat{\phi}(\sigma/(\sqrt{2}b)) \lesssim b^{\alpha+1}$$

because $\|\hat{K}\|_\infty \le \|K\|_1 < \infty$, the upper tail of a Gaussian distribution is bounded above by

$$\int\limits_{1/b}^{\infty} \hat{\phi}(\sigma t)\,\mathrm{d}t \lesssim \frac{b}{\sigma^2}\hat{\phi}(\sigma/b)$$

and $(\sigma/b)^2 = O(\log(1/b^{\alpha+1}))$ by assumption. The assertion follows. □

References

1. Al Labadi, L., Abdelrazeq, I.: On functional central limit theorems of Bayesian nonparametric priors. Stat. Methods Appl. **26**(2), 215–229 (2017)
2. Al Labadi, L., Zarepour, M.: On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. Bayesian Anal. **8**(3), 553–568 (2013)
3. Conti, P.L.: Approximated inference for the quantile function via Dirichlet processes. Metron **62**(2), 201–222 (2004)
4. Dattner, I., Goldenshluger, A., Juditsky, A.: On deconvolution of distribution functions. Ann. Statist. **39**(5), 2477–2501 (2011)
5. Dattner, I., Reiß, M., Trabs, M.: Adaptive quantile estimation in deconvolution with unknown error distribution. Bernoulli **22**(1), 143–192 (2016)
6. Dedecker, J., Fischer, A., Michel, B.: Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. Electron. J. Statist. **9**(1), 234–265 (2015)
7. Fan, J.: On the optimal rates of convergence for nonparametric deconvolution problems. Ann. Statist. **19**(3), 1257–1272 (1991)
8. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Statist. **1**, 209–230 (1973)
9. Ghosal, S., van der Vaart, A.: Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 44. Cambridge University Press, Cambridge (2017)
10. Giné, E., Nickl, R.: Rates of contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. Ann. Statist. **39**(6), 2883–2911 (2011)
11. Giné, E., Nickl, R.: Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 40. Cambridge University Press, New York (2016)
12. Hall, P., Lahiri, S.N.: Estimation of distributions, moments and quantiles in deconvolution problems. Ann. Statist. **36**(5), 2110–2134 (2008)
13. Hjort, N.L., Petrone, S.: Nonparametric quantile inference using Dirichlet processes. In: Advances in Statistical Modeling and Inference. Ser. Biostatistics, vol. 3, pp. 463–492. World Sci. Publ., Hackensack, NJ (2007)

14. Meister, A.: Deconvolution Problems in Nonparametric Statistics. Lecture Notes in Statistics, vol. 193. Springer, Berlin (2009)
15. Rousseau, J., Scricciolo, C.: Wasserstein Convergence in Bayesian Deconvolution Models. https://arxiv.org/abs/2111.06846 (2021)
16. Scricciolo, C.: Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. Bayesian Anal. **9**(2), 475–520 (2014)
17. Scricciolo, C.: Bayesian Kantorovich deconvolution in finite mixture models. In: New Statistical Developments in Data Science. Springer Proceedings in Mathematics & Statistics, vol. 288, pp. 119–134. Springer, Cham (2019)

Can the Compositional Nature of Compositional Data Be Ignored by Using Deep Learning Approaches?

Matthias Templ

Abstract The application of non-compositional methods to compositional data representing parts of a whole should be avoided from a theoretical point of view. For example, one can show that Pearson correlations applied to compositional data are biased to be negative. Moreover, almost all statistical methods lead to biased estimates when applied to compositional data. One way out is to analyze data after representing them in log-ratio coordinates. However, several implications arise, such as interpretation in log-ratios and dealing with zeros and non-detects where log-ratios are undefined. When focusing on settings where only the prediction and classification error is important rather than an interpretation of results, one might argue to rather use non-linear methods to avoid those implications. Generally, it is known that misclassification and prediction errors are lower with a log-ratio approach when using machine learning methods that model the linear relationship between variables. However, is this also true when training a neural network who may learn the inner relationships between parts of a whole also without representing the data in log-ratios? This paper gives an answer of this matter based on applications with multiple real data sets, leading to the recommendation to use a compositional treatment of compositional data in any case. Misclassification and prediction errors are lower when nonlinear methods, and in particular deep learning methods, are applied together with a compositional treatment of compositional data. The compositional treatment of compositional data therefore remains very important even in the context of focusing on prediction errors using deep artificial neural networks or other nonlinear methods.

M. Templ (✉)
Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Rosenstrasse 3, 8404 Winterthur, Switzerland
e-mail: matthias.templ@zhaw.ch

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_11

1 Introduction

1.1 Compositional Data

Composite data appear to be multivariate observations whose parts describe a whole. Strictly speaking, every single value is a real number greater than zero. For compositional data, only the relative information of the observations is of interest. Let us consider a D-part composition $x = [x_1, \ldots, x_D]'$ with strictly positive parts $x_1, ..., x_D$. The same relative information is contained in x_i/x_j and $(ax_i)/(ax_j)$ for any non-zero scalar value a. The composition can be re-expressed as proportions, $x^* = ax$ by setting $a = 1/\sum x_i$, but the approximation to 1 is arbitrary and not relevant [14], since the log-ratios remain unchanged. The composition x^* then belongs to the standard simplex defined by

$$\left\{ x^* = \left[x_1^* \ldots x_D^* \right]' \mid x_i^* > 0, \sum_{i=1}^{D} x_i^* = 1 \right\} \quad .$$

The simplex geometry is different from the usual Euclidean geometry for which our statistical tools were developed. However, we can define operations on the simplex. Even more, a norm, an inner product and a distance can be defined suitable for the simplex vector space.

Consider two compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ and $\mathbf{y} = (y_1, \ldots, y_D)'$ of S^D. An Euclidean linear vector space structure (*Aitchison geometry*) on S^D is defined by [2]. The following operations replace addition and multiplication in the traditional sense, with perturbation $\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, x_2 y_2, \ldots, x_D y_D)'$ and powering by a constant $\alpha \in \mathbb{R}$, $\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)'$. The Aitchison inner product is defined by $\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$. The Aitchison norm is $\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}$. Finally, a distance can be formulated—the Aitchison distance—given by $d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}$.

1.2 Expressing Compositional Data in Log-Ratio Coordinates

While this notation of distances and mathematical operations is helpful when working with compositional data itself, a different strategy is required when applying existing statistical methods to compositional data. In order to apply standard methods of multivariate data analysis, the compositions must be represented in Euclidean space, e.g., by expressing the information in log-ratio coordinates. This is typically achieved by using isometric log-ratio transformations introduced in Egozcue et al. [8].

First, we introduce the representation of compositions in centered log-ratio coordinates. The relative information of the part x_1 can be expressed as: $\ln \frac{x_1}{x_2}, \ln \frac{x_1}{x_3}, \ldots, \ln \frac{x_1}{x_D}$. This can be "aggregated" to $y_1 = \frac{1}{D}\left(\ln \frac{x_1}{x_2} + \cdots + \ln \frac{x_1}{x_D}\right) = \ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}$. y_1 is called the centered log-ratio (clr) coordinate [14]. However, there are other log-ratios, namely all pairwise logarithms.
$\ln \frac{x_1}{x_2}, \ldots, \ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_3}, \ldots, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}$. They can in turn be *aggregated* into clr coordinates by

$$\mathbf{y} = \mathrm{clr}(\mathbf{x}) = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right)' . \quad (1)$$

Note that the resulting centered log-ratio coordinates are singular, since $y_1 + \cdots + y_D = 0$. The clr coordinates thus form the composition from the simplex S^D onto a $(D-1)$-dimensional hyperplane in $\mathbb{R}^D$.

While the centered log-ratio transformation is relatively simple to explain and thus often used in practice, there are some disadvantages, especially for data sets with many variables. The denominator for high-dimensional compositional data usually hardly shows any data structure, but rather only noise. This means that the main role in the observations is then played by the dominance of the components in the log ratios, and, from a more technical point of view, the log-ratio of the geometric means in the comparison of the log and Aitchison distances is almost 0. Let d_e^2 the Euclidean distance and d_a^2 the Aitchison distance, we get [22]

$$\begin{aligned}
\mathrm{d}_e^2(\log(\mathbf{x}), \log(\mathbf{y})) &= \sum_i (\log(x_i) - \log(y_i))^2 \\
&= \sum_i \left(\log \frac{x_i}{\mathrm{g_m}(\mathbf{x})} - \log \frac{y_i}{\mathrm{g_m}(\mathbf{y})} + \log \frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})} \right)^2 \\
&= \sum_i \left(\log \frac{p_i}{\mathrm{g_m}(\mathbf{p})} \right)^2 + 2 \log \frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})} \sum_i \left(\log \frac{p_i}{\mathrm{g_m}(\mathbf{p})} \right) + D \log^2 \left(\frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})} \right) \\
&= \mathrm{d}_a^2(\mathbf{x}, \mathbf{y}) + D \log^2 \left(\frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})} \right) \geq \mathrm{d}_a^2(\mathbf{x}, \mathbf{y})
\end{aligned} \quad (2)$$

Note that $\mathbf{p}$ is the perturbation difference $\mathbf{p} = \mathbf{x} \ominus \mathbf{y} = (x_1/y_1, x_2/y_2, \ldots, x_D/y_D)$. The clr works with geometric means, and this means—see also Eq. (2)—that the log and centred log-ratio solutions gets more similar the higher the number of parts/variables in a data set.

The class of isometric log-ratio (ilr) coordinates aims to form an orthonormal basis in a hyperplane and express the composition in it. The resulting vector $\mathbf{z}$ is in $\mathbb{R}^{D-1}$, i.e. the isometric log-ratio transformation maps the data from the simplex to the Euclidean vector space.

One particular choice of a basis leads to

$$\text{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$$

with

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, \quad \text{for} \quad j = 1, \ldots, D-1. \tag{3}$$

These ilr coordinates are referred as *pivot* (log-ratio) *coordinates* [14]. Such a choice has also a primary importance for the coordinate system as a whole. z_j summarizes now all relative information (log-ratios) about x_j, and can thus be interpreted as the relative dominance of x_j within the given composition. These special pivot coordinates are used for model-based imputation of missing values or rounded zeros, but also generally often in a regression context. Finally, $z_j = 0$ indicates a balanced state between x_j and an average behavior of the other parts in the given composition.

Isometric log-ratio (ilr) coordinates $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ form an orthonormal basis [8]. Even more, both clr-coordinates and ilr-coordinates represent an isometry. For two compositions $\mathbf{x}$ and $\mathbf{y} \in S^D$ and $c \in \mathbb{R}$ it holds that $\text{ilr}(\mathbf{x} \oplus \mathbf{y}) = \text{ilr}(\mathbf{x}) + \text{ilr}(\mathbf{y})$, $\text{ilr}(c \odot \mathbf{x}) = c \cdot \text{ilr}(\mathbf{x})$, $\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y}) \rangle$, $\|\mathbf{x}\|_a = \|\text{ilr}(\mathbf{x})\|$ and $d_a(\mathbf{x}, \mathbf{y}) = d(\text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y}))$ [see also 14, for example]. Thus, all metric concepts in the simplex are preserved when using ilr coordinates.

In addition, for several applications also all possible log-ratios may be used instead of centered or isometric log-ratios [12]. That is, each variable is divided by one of the other variables before the logarithm is taken. For a data set with 10 variables, there are already 45 possible log ratios between the variables. Malyjurek et al. [25] suggested using all possible log-ratios. Since this leads to a large number of log-ratios, they suggest feature selection to reduce the number of log-ratios obtained. They argue that a centered log-ratio transformation may *average too much* leading to a higher false discovery rates of biomarkers [25].

1.3 Requirements for Compositional Data Analysis and Difficulties in Practice

The first requirement is scale invariance. This means that the same results are obtained when a composition is multiplied by a constant. Another requirement is permutation invariance. This is achieved when a reordering of parts (variables) leads to the same results. Subcompositional coherence is another requirement and it can be divided into two categories, namely dominance and ratio preservation. Dominance means that the distance between two compositions must be greater

than the distance between a sub-composition of these two compositions. The ratio preserving condition is fulfilled when unselected parts have no influence on the results.

Outlook:
In the next section, a critism and implications related to a compositional analysis are raised. As a potential way out to avoid those implications, deep artificial neural networks will be introduced in Sect. 3. Various classification and prediction methods are compared to a bunch of data sets and different data pre-processing and classification methods that are introduced or mentioned in Sect. 4. Section 5 contains the results and the interpretation of these results. The last section summarizes and discusses whether a compositional treatment of compositional data is unnecessary when non-linear methods such as deep artificial neural networks are used.

2 Problems and Difficulties in Practice

The reasons why practitioners often dislike working with compositional data analysis are as follows:

- For a compositional analysis, statistical methods are usually applied to log-ratio coordinates. If this is the case, then the interpretation should not be done on the original scale of the data, but must also be done in log-ratios, which is a major challenge.
- Finding so called balances [7] for proper and useful isometric log-ratios is sometimes difficult and time-consuming.
- Small values are outliers in the simplex. This is also true when expressed in log-ratio coordinates. Such small values often occur in composition data, for example, when the chemical concentration of an element is low.
- In practice, many data sets contain true zeros, rounded zeros or missing values. A log-ratio with a 0 in the nominator or denominator is undefinded, and replacing zeros is non-trivial [39] and somewhat artificial in the case of real zeros [40]. Examples:
  - Geochemical data: values below detection limit of a measurement unit. These values must be properly replaced within zero and the corresponding detection limit of a measurement unit per variable.
  - Microbiome data includes a lot of zero counts. The issue is then to replace these zeros with values in (0, 1], for example.
  - Compositional parts in official statistics, e.g. on income, may include zeros but also missing values from non-responses. Before a log-ratio analysis, the missings must be imputed and the zeros must be handled in some way.

Various authors criticized the strict use of the principles of a compositional data analysis [30, 31] and applied other concepts violating the principles to avoid some of those difficulties [4, 21, 32, 35, 41].

For many problems, one just want to keep the prediction quality as high as possible. If the quality of the prediction is more important than the interpretation, we can reconsider and ask ourselves whether all the requirements for an appropriate data composition analysis method must be met, which often complicates a compositional analysis.

In linear methods, it is known that the prediction quality is higher when the composition of the data is taken into account than when linear methods are used without this consideration [see, e.g., 13]. Thus after representing compositional data in log-ratio coordinates, we can exect better predictive power of methods that relies on the linear relationship between variables.

The question is if instead of log-ratio methods with all their difficulties and instead of forging new complex theories to better deal with zeros and missing values, we simply try to use non-linear methods and answer the question whether one can (mostly) neglect the compositional nature of compositional data for classification and prediction problems. Can deep neural networks learn the dependencies in compositional data? Even if the methods rely on Euclidean geometry, but compositions are restricted in a simplex sample space?

3 Artificial Neural Networks

Neural networks are well implemented today, and recent advances, e.g., in TensorFlow [1, 3] and keras [5], have better weighting and activation functions for the neural layers, better optimization techniques, e.g., Adam [19], and are more efficient than implementations of neural networks in the past, allowing more layers ($\sim$*deep*) to be used. In the following, we will always refer to a deep artificial neural network as an artificial neural network for simplicity.

A neural network is just a non-linear statistical model [15], which is based on a transformed (with an activation function) weighted linear combinations of the sample values. Each neuron is formed by transformed weights and thus a neuron hold some information on each variable and a set of observations. Each neuron has an activation, depending on how the input information looks like. A layer in a deep neural network is a collection of neurons in one step. There are three types of layers: the input layer, the hidden layers, and the output layer. The goal of training a network is to find the optimal weights for each connection between the neurons of two layers. Initially all weights are chosen randomly, and a loss function of the network is selected. The output of such a loss function is a single number judging the quality of the neural network. To lower the value of the loss function, an adaptive moment estimation called Adam [19, 28] is used (other methods can be selected), which is a stochastic gradient descent method that uses adaptive learning rates for each parameter of the algorithm. With this gradient descent optimization all the weights are optimized to

reach the next local minimum resulting in the most rapid decrease of the loss function. The (stochastic) gradient is used to adjust these weights in each step (epoch) using back propagation, whereby the weights gets updated. Choosing a proper activation function, a deep neural network is able to find non-linear relationships between a target variable and predictors.

Next to the choice of the activation function (here we used reLu [16]) a lot of hyper-parameters must be chosen. In brackets is our choice: the loss (mean squared error) and evaluation (absolute error) metrics, the number of epochs (500), a patience value (50), the number of layers (10), the number of neurons in each layer (1000, 900, ..., 100), possible dropout (10% in the first 5 layers), ratio of validation data (20%) are the most important hyper-parameters to choose adequately to prevent over- or underfitting.

It should be noted that in this work we do not consider more general multilayer perceptron networks with convolutional layers or recurrent layers.

4 Data Sets, Their Treatment and Selected Methods

First, it should be pointed out that a model-based simulation study with artificially generated data is not very useful, because the results of non-compositional methods are better if the response is generated with non-compositional methods. If the response is generated with the log-ratio coordinates, the compositional methods will perform better. Moreover, real data in the classification and regression context correspond much closer to the truth than results in a simulation. Therefore, no simulation was performed.

To demonstrate if one can ignore the compositional nature in a classification context some real-world data sets.

Aging of beers: The 48 beers of the beer data set [42] had been measured before and after beer aging, resulting in 96 measurements on 17 variables using a gas spectrometer. It should be noted that beer experts can recognize a beer brand by its consistent, fresh flavor. Aging of beer leads to the formation of stale flavors, regardless of whether it is heat-related or insufficient storage; thus, the original flavor is lost. A good classifier should be able to tell if a measurement is from an aged beer, because the chemical composition differs [37].

Aging of honey: This data set consists of 429 observations and 17 variables. Chemical profiling of honey [29, 43] is an important issue when determining the botanical and geographical origins of them. Honey consists mainly of sugars and water as well as smaller components such as minerals, vitamins, amino acids, organic acids, flavonoids and other phenolic compounds and flavourings [6, 29]. The flowers, the geographical regions, the climate and the species of bees involved in the production are the main factors that determine the composition, colour, aroma and taste of the honey [6, 10]. As mislabelling and adulteration of honey has unfortunately become a worldwide problem, it is not only crucial to detect

the adulterations in honey and thus to classify new honey samples correctly. The problem of the performance of techniques for detecting adulteration in honey is widely discussed [11, 33, 34, 38]. The target variable is the type of the "honey" (honey, adulterated honey, sirup).

Coffee: 30 commercially available coffee samples of different origins have been analysed. In the original data set, 15 volatile compounds (descriptors of coffee aroma) were selected for a statistical analysis. As in Korhonová et al. [20] we selected six compounds (compositional parts) on three sorts of coffee. The labels of these three sorts are used as target variable.

Yatquat fruit: 40 observations on the yatquat tree fruit on 7 variables [2] represent the quality of yatquat tree fruit in terms of the relative volume proportions of flesh, skin and stone. The yatquat tree produces a single large fruit each season. The data include the fruit compositions of the current season, the compositions of the fruit of the same 40 trees in the previous season when none of the trees were treated, and additionally the type: 1 for the treated trees, -1 for the untreated trees. The type is used as target variable.

Agricultural land soil: Geochemical data set on agricultural and grazing land soil with 2108 observations and 30 variables [26]. The sampling, at a density of 1 site/2500 km^2, was completed at the beginning of 2009 by collecting 2211 samples of agricultural soil (Ap-horizon, 0–20 cm, regularly ploughed fields), and 2118 samples from land under permanent grass cover (grazing land soil, 0–10 cm), according to an agreed field protocol. All GEMAS project samples were shipped to Slovakia for sample preparation, where they were air dried, sieved to <2 mm using a nylon screen, homogenised and split to subsamples for analysis. They were analysed for a large number of chemical elements. In this sample, the main elements by X-ray fluorescence are included as well as the composition on sand, silt and clay. Mean temperature and annual precipitation served as target variables.

Eight different ways to pre-process the data set are considered, namely

- No transformation (original non-normalized data). In the graphics noted as *no*. Some variables with larger values may dominate the result.
- Normalised data by subtracting the arithmetic mean and dividing by the standard deviation (applied for each variable), thus each variable is normalized to mean zero und variance 1. Noted in the figures as *normalized*. This would make the contribution of each variable comparable in amount.
- Logarithmic transformation applied on each part, i.e. the logarithm of values are used. Noted in the figures as *log*. This *symmetrize* the typicall right-skewness of variables.
- Logarithmic transformation followed by normalization to mean 0 and variance 1 (*norm* + *log*). This is a symmetrisation plus normalisation so that each variable has approximately the same influence.
- Data expressed as percentages (*percentages*). Often practitioner express the data in percentages or row-wise sum to 1 or another constant.
- Data expressed as percentages + normalized (*percentages norm*). As before, but the influence of each variable is *averaged*.

- Centered log-ratio coordinates (*centered coordinates*). The normalization is done in a compositional way. This also symmetrise the variables, see Eq. 1.
- Pivot log-ratio coordinates (*pivot coordinates*). The normalization is done in a compositional way. These coordinates are orthogonal to each other. See Eq. 3.

All in all six classifiers are compared. Namely,

- Naive Bayes: As a representative of a kernel and Bayesian method. It is a classification technique based on Bayes' theorem, assuming independence of predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a certain feature in a class is not related to the presence of another feature. They are among the simplest Bayesian network models, but can achieve higher degrees of accuracy when combined with kernel density estimation. We use the implementation in Majka [24].
- Linear discriminant analysis (lda): As a representative of a classical discrimination linear method based on covariance estimation. Linear discrimination tries to find linear functions which separate the observations into different classes. Observations within one group should have as many similar features as possible whereas observations of different groups should have little in common. The decision boundaries between classes are used to classify new observations.
- k nearest neighbor classification (knn): As a representative of a simple classification model that do not rely on parameter distribution assumptions. An observation is classified by a plurality vote of its k neighbors, with the class of a new observation being assigned to the class most common among its k nearest neighbors. The approach purely depends on distances, and represents a simple but sometimes quite effective classification method.
- Generalized linear models (glm): As a representative of a linear model. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function. Thus, the target variable need not be continuous, but can also be treated as binomial or multinomial.
- Random forest (cforest): As a representative of a non-linear method based on classification trees. A random forest consists of a large number of individual decision trees whereby the number and the predictors itself are randomly chosen and the observations are bootstrapped.
- (Deep) Artificial neural networks (ann): As a representative of a non-linear method in deep learning. See Chap. 2 for details.

Thus 48 results are compared for each data set.

The correct classification of elements in the target variable is an essential element of any study. Misclassification occurs when individuals are assigned to a category other than the one they should belong to. The misclassification rates, the percentage of incorrectly classified instances, are estimated in a 10 times repeated 10-fold cross validation setting. However, for the artificial neural network a cut-off was made due to high computation times and a simplified cross validation with 5 times repeated random assignment in 75% training data and the rest as test data was used and the average reported. The purpose of this different setting is not to compare the methods,

but to evaluate the effects of preprocessing the data for each method separately with as good choice of estimating the prediction errors and accuracy as possible.

In addition to the classification context, we also study the outcome in a regression context. The mean absolute error, $|\hat{\mathbf{y}} - \mathbf{y}|$, evaluated on test data in a cross-validation setting is one representative to express the prediction error.

5 Results

Figure 1 represents an extended analysis of Templ and Templ [37] using artificial neural networks.

It is shown that a classifier applied to the pivot coordinates gives in principle better results, i.e. the analysis of the data composition leads to lower misclassification rates than a pure normalization or logarithmization of variables. Random forests and artificial neural networks clearly represent non-linear methods. While a pivot log-ratio transformation lowers the misclassification rate with random forests, the results for logarithmised and normalised data are best with with artificial neural networks, followed by pivot coordinates. The presentation in constant row sums (e.g. as percentages) have negative influence on almost all (non-compositional) methods,

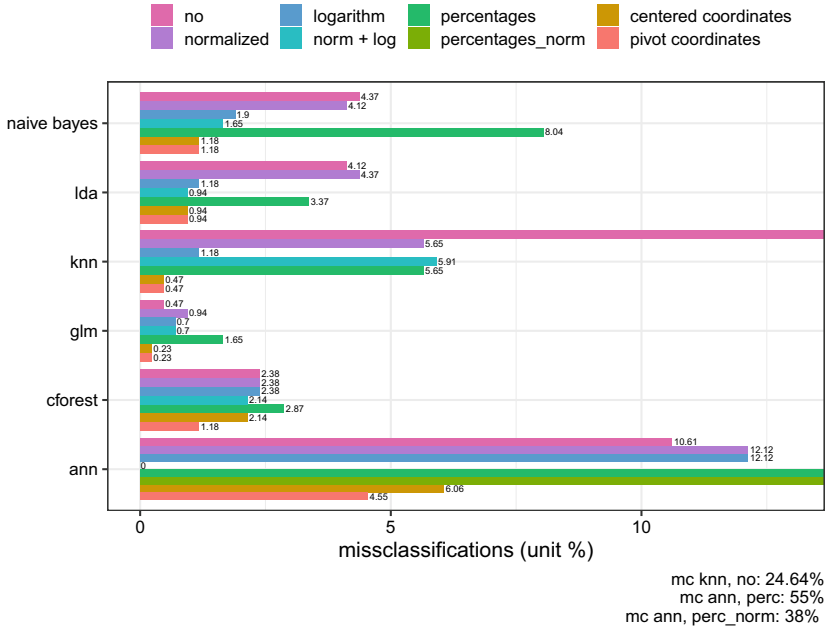


Fig. 1 Misclassification rate (in %) for different methods, transformations and normalizations, for the beer dataset

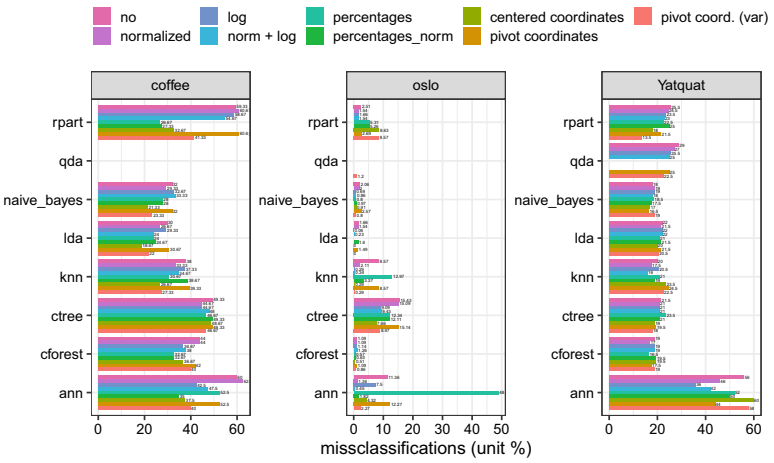


Fig. 2 Misclassification rate (in %) for different methods, transformations and normalizations, for the coffee, oslo and yatquat data sets

but especially the artifical neural network cannot deal with such constraints and results to 55% misclassification rates.

The results for the Yatquat data, see Fig. 2 shows generally lower misclassification rates (and thus higher accuracy) when compositional approaches are used to treat the data set. A compositional *normalization* is thus preferable compared to a normalization to mean 0 and variance 1. The artificial neural network does not provide good solutions at all. The reason is simple, the data set is just too small for an artificial neural network. One can reduce the number of layers and neurons, but the results stays comparable. Generally, a compositional treatment of the Oslo data leads to better results, which is also true for the coffee data set.

The performance of the methods are also investigated in a regression context using the gemas data set. The models may not be the most favourable in practice: who wants to predict the mean temperature or annual precipitation from the soil composition. Rather one is interested in the opposite: does the precipitation and temperature in a region defines the sand, silc and clay composition, for example. However, with this choice we get some insights into the performance of the different methods (Fig. 3).

The annual precipitation as well as the mean temperature are once modelled by the chemical elements Al, Ba, Ca, Fe, K, Mg, Mn, Na, Nb, P, Si, Sr, Ti, V, Y, Zn and Zr and once modelled with the values on sand, silt and clay. The results are comparable. Centered log-ratio coordinates are singular as well as the representation in percentages. Robust methods run into problems for singular data, see some results from robust MM regression (*rlm*). One reason for the similarity of the results is the weaknesses of the models. The annual precipitation cannot be explained by the composition of sand, silt and clay. Also the composition of chemical elements provides only a weak performance. The chemical elements and the composition of sand, silt and clay result in a model that provides some explanations for the mean

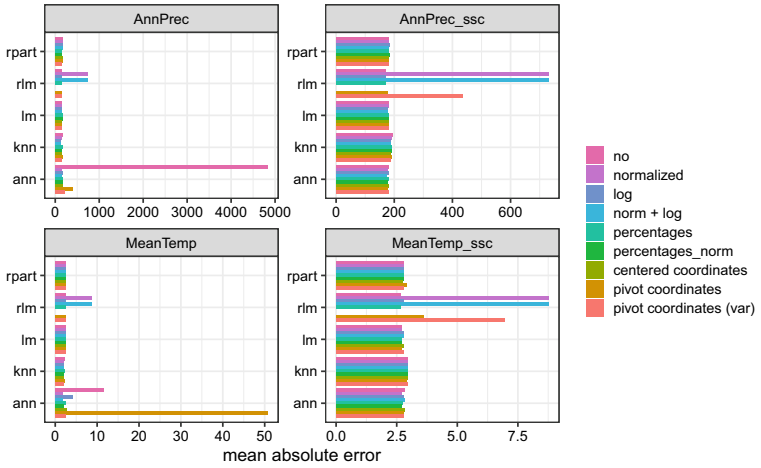


Fig. 3 Prediction error for different methods, transformations and normalizations, for the gemas data set

temperature. For example, the R^2 of the linear classical least squares model is 0.48. However, there are no major differences between the regression methods and between the normalisation and transformation methods.

6 Discussion and Conclusion

Templ [36] already studied the performance of imputing rounded zeros with artificial neural networks (similar context with and without considering composition), and Lubbe et al. [23] studied the performance of imputing rounded zeros from high-dimensional bionomic data. Both showed that a compositional treatment of data is important even for artificial neural networks and improves the results for the imputation of zeros. This paper extends these results with respect to classification and regression problems.

In this contribution, we showed that not only for linear methods (such as linear discriminant analysis or generalized linear models), but also distribution-free methods (such as k nearest neighbor classification) and non-linear methods like random forests results improves in general when a compositional treatment of data is applied. This is also true for neural network. They learn better resulting in higher predictive power when the data is represented in pivot coordinates. Applying a neural network to data without considering the special nature of compositional data leads to worse results. For example, the artificial neural network does not learn all non-linear (compositional) relationships as well as when the data is represented in log-ratio coordinates, even though enough layers and neurons were chosen. A hyperparameter setting was used that was learned in preliminary studies by trying different settings.

Here, the training, validation, and testing errors were comparable, so overfitting is not expected.

For the classification of the real-world data sets, the results are not black/white but a tendency is visible. Based on our results, we recommend a compositional pretreatment of the data before applying any machine learning or deep learning method. The tendency is that this provides better results and even the neural network has an easier time learning the non-linear (compositional) relationships. And it can go really wrong when working with untransformed representation of the data in percentages (or generally with data having constant row-sums). For classification problems where a misclassification rate is in the foreground, a compositional approach is therefore also advantageous for an artificial neural network. The results from the regression analysis wasn't that clear. It would need further work using other data sets to also judge for the regression context.

The main finding is that—although an artificial deep neural network can learn non-linear dependencies—by taking into account the compositional nature of the data by using appropriate logarithmic representations of the data is beneficial also before applying such non-linear methods. The results can improve significantly and the stochastic gradient algorithms used to optimise the network perform better in log-ratio coordinates. This results are not surprising, since a artificial neural network also works in the Euclidean geometry, i.e. all distances used are of Euclidean (or Manhattan) nature. So a representation in log-ratio coordinates, i.e. a representation in $\mathbb{R}^{D-1}$ instead of in simplex, helps the network to learn faster and better. Those who believed that the use of a non-linear method made a compositional treatment of compositional data unnecessary is proven to be wrong.

If the number of variables/parts increases, the results of a log transformation together with a traditional normalization to mean 0 and variance 1 gives comparable results obtained with compositional approaches. The reason is already expressed in Eq. 2. For example, the oslo data set contains a lot of parts, and the results do not show large differences between log and log-ratio transformation anymore. One could then come up with the idea of using the simpler concept of a logarithmic transformation instead of a compositional approach. But the use of a log-ratio transformation has further advantages. Through this, data sets are automatically normalised, e.g. the centred log-ratio transformation normalises each value of a composition with the geometric mean value of the composition. Especially in the microbionom and omics sciences, the use of complex instruments such as spectrometers and their calibration is difficult. One sample may have a different baseline than another sample, so normalisation is required to compare the samples. Different types of normalisation then lead to different results. This is not the case with composition normalisation and the samples can be compared as only the ratios between the parts are analysed.

A pivot coordinate representation gives slightly better results than using centered log-ratio coordinates in general. Future work includes the influence of weighted pivot coordinates [18] on the results of non-linear classification methods.

An alternative to using log-ratio transformations before applying an artificial neural network would be to replace the Euclidean distances used internally in TensorFlow with Aitchison distances, so that any internal distance computation and the evalu-

ation function works with Aitchison distances. The loss function would then also have to be chosen differently to meet the requirements of a compositional analysis. However, such changes to TensorFlow are beyond the scope of this paper, and we have shown in Sect. 1.2 that the Euclidean distances after an isometric transformation are equal to the Aitchison distances.

Future work includes testing the methods in a regression context on various other data sets and work with all possible log-ratios, i.e. not only with privot or centered log-ratio coordinates, using regularization techniques when the number of parts is moderately low.

The hyperparameters of the applied neural networks was chosen based on an evaluation and comparison of the training, validation and test errors, and also stop rules are applied if a higher number of epochs does not improve the value of the loss function at 50 more epochs. The selection of the values of the hyperparameters can be extended by searching for an optimal neural architecture by optimizing its hyperparameters. Active research is being conducted in this area, and currently problems and data in computer vision and natural language processing are being addressed using these new approaches, see Ren et al. [9], Escalante [17], Wistuba et al. [27], He et al. [44] for more information.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C.: Tensorflow: large-scale machine learning on heterogeneous systems (2018). https://www.tensorflow.org/. Version: 1.10.0, Software available from tensorflow.org
2. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (1986)
3. Allaire, J.J., Tang, Y.: Tensorflow: R Interface to 'TensorFlow' (2019). https://github.com/rstudio/tensorflow. R package version 2.0.0
4. Butler, A., Glasbey, C.: A latent gaussian model for compositional data with zeros. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **57**(5), 505–520 (2008). https://doi.org/10.1111/j.1467-9876.2008.00627.x
5. Chollet, F., et al.: Keras (2015). https://keras.io
6. da Silva, P.M., Gauche, C., Gonzaga, L.V., Costa, A.C.O., Fett, R.: Honey: chemical composition, stability and authenticity. Food Chem. **196**, 309–323 (2016). ISSN 0308-8146. https://doi.org/10.1016/j.foodchem.2015.09.051
7. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**(7), 795–828 (2005)
8. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geol. **35**(3), 279–300 (2003)
9. Escalante, H.J.: Automated Machine Learning—A Brief Review at the End of the Early Years, pp. 11–28. Springer International Publishing, Cham (2021)
10. Escuredo, O., Dobre, I., Fernández-González M., Seijo, M.C.: Contribution of botanical origin and sugar composition of honeys on the crystallization phenomenon. Food Chem. **149**, 84–90 (2014). ISSN 0308-8146. https://doi.org/10.1016/j.foodchem.2013.10.097
11. Fakhlaei, R., Selamat, J., Khatib, A., Faizal, A., Razis, A., Sukor, R., Ahmad, S., Babadi, A.A.: The toxic impact of honey adulteration: a review. Foods **9**(11) (2020). ISSN 2304-8158. https://doi.org/10.3390/foods9111538

12. Filzmoser, P., Walczak, B.: What can go wrong at the data normalization step for identification of biomarkers? J. Chromatogr. A **1362**, 194–205 (2014). ISSN 0021-9673. https://doi.org/10.1016/j.chroma.2014.08.050
13. Filzmoser, P., Hron, K., Templ, M.: Discriminant analysis for compositional data and robust estimation. J. Comput. Stat. **27**(4), 585–604 (2012)
14. Filzmoser, P., Hron, K., Templ, M.: Applied Compositional Data Analysis. Springer International Publishing (2018). ISBN 9783319964225. https://doi.org/10.1007/978-3-319-96422-5
15. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009). ISBN 978-0-387-84857-0
16. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification (2015)
17. He, X., Zhao, K., Chu, X.: AutoML: a survey of the state-of-the-art. Knowl. Based Syst. **212**, 106622 (2021). ISSN 0950-7051. https://doi.org/10.1016/j.knosys.2020.106622
18. Hron, K., Menafoglio, A., Palarea-Albaladejo, J., Filzmoser, P., Talská, R., Egozcue, J.J.: Weighting of parts in compositional data analysis: advances and applications. Math. Geosci. **54**, 71–93 (2022). https://doi.org/10.1007/s11004-021-09952-y
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR. abs/1412.6980 (2014)
20. Korhonová, M., Hron, K., Klimcíková, D., Müller, L., Bednář, P., Barták, P.: Coffee aroma-statistical analysis of compositional data. Talanta **80**, 710–715 (2009). https://doi.org/10.1016/j.talanta.2009.07.054
21. Leininger, T.J., Gelfand, A.E., Allen, J.M., Silander, J.A.: Spatial regression modeling for compositional data with many zeros. J. Agric. Biol. Environ. Stat. **18**(3), 314–334 (2013). https://doi.org/10.1007/s13253-013-0145-y
22. Lovell, D., Müller, W., Taylor, J., Zwart, A., Helliwell, C.: Proportions, percentages, PPM: do the molecular biosciences treat compositional data right? In: Compositional Data Analysis: Theory and Applications, pp. 191–207. Wiley (2011). https://doi.org/10.1002/9781119976462.ch14
23. Lubbe, S., Templ, M., Filzmoser, P.: Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemom. Intell. Lab. Syst. **215**, 104248 (2021)
24. Majka, M.: Naivebayes: high performance implementation of the Naive Bayes algorithm in R (2019). https://CRAN.R-project.org/package=naivebayes. R package version 0.9.7
25. Malyjurek, Z., de Beer, D., Joubert, E., Walczak, B.: Working with log-ratios. Anal. Chimica Acta **1059**, 16–27 (2019). ISSN 0003-2670. https://doi.org/10.1016/j.aca.2019.01.041
26. Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., Akinfiev, G., Albanese, S., Amashukeli, Y., Andersson, M., Arnoldussen, A., Artamonov, Y., Audion, A., Baritz, R., Barker, K., Batista, M., Bellan, A., Belougushev, V., Bitz, I., Branellec, M., Zomeni, Z.: Chemistry of Europe's Agricultural Soils—Part A: Methodology and Interpretation of the Gemas Data Set (2014). ISBN 978-3-510-96846-6
27. Ren, P., Xiao, Y., Chang, X., Huang, P., Li, Z., Chen, X., Wang, X.: A comprehensive survey of neural architecture search: challenges and solutions. ACM Comput. Surv. **54**(4). ISSN 0360-0300. https://doi.org/10.1145/3447582
28. Ruder, S.: An overview of gradient descent optimization algorithms (2016). arXiv: 1609.04747
29. Santos-Buelga, C., González-Paramás, A.M.: Chemical Composition of Honey, pp. 43–82. Springer International Publishing, Cham (2017). ISBN 978-3-319-59689-1
30. Scealy, J.L., Welsh, A.H.: Regression for compositional data by using distributions defined on the hypersphere. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **73**(3), 351–375 (2011). https://doi.org/10.1111/j.1467-9868.2010.00766.x
31. Scealy, J.L., Welsh, A.H.: Colours and cocktails: compositional data analysis 2013 lancaster lecture. Aust. NZ J. Stat. **56**(2), 145–169 (2014). https://doi.org/10.1111/anzs.12073
32. Scealy, J.L., Wood, A.T.A.: Score matching for compositional distributions (2020)
33. Se, K.W., Wahab, R.A., Syed Yaacob, S.N., Ghoshal, S.K.: Detection techniques for adulterants in honey: challenges and recent trends. J. Food Compos. Anal. **80**, 16–32 (2019). ISSN 0889-1575. https://doi.org/10.1016/j.jfca.2019.04.001

34. Soares, S., Amaral, J.S., Oliveira, M.B.P.P., Mafra, I.: A comprehensive review on the main honey authentication issues: production and origin. Compr. Rev. Food Sci. Food Saf. **16**(5), 1072–1100 (2017). https://doi.org/10.1111/1541-4337.12278
35. Stewart, C., Field, C.: Managing the essential zeros in quantitative fatty acid signature analysis. J. Agric. Biol. Environ. Stat. **16**(1), 45–69 (2011). https://doi.org/10.1007/s13253-010-0040-8. March
36. Templ, M.: Artificial Neural Networks to Impute Rounded Zeros in Compositional Data, pp. 163–187. Springer International Publishing, Cham (2021). ISBN 978-3-030-71175-7
37. Templ, M., Templ, B.: Analysis of chemical compounds in beverages—guidance for establishing a compositional analysis. Food Chem. **325**, 1–7 (2020)
38. Templ, M., Templ, B.: Statistical analysis of chemical element compositions in food science: problems and possibilities. Molecules **26**(19) (2021). https://doi.org/10.3390/molecules26195752
39. Templ, M., Hron, K., Filzmoser, P., Gardlo, A.: Imputation of rounded zeros for high-dimensional compositional data. Chemometr. Intell. Lab. Syst. **155**, 183–190 (2016). https://doi.org/10.1016/j.chemolab.2016.04.011
40. Templ, M., Hron, K., Filzmoser, P.: Exploratory tools for outlier detection in compositional data with structural zeros. J. Appl. Stat. **44**(4), 734–752 (2017). https://doi.org/10.1080/02664763.2016.1182135
41. Tsagris, M., Stewart, C.: A folded model for compositional data analysis. Aust. NZ J. Stat. **62**(2), 249–277 (2020). https://doi.org/10.1111/anzs.12289
42. Varmuza, K., Steiner, I., Glinsner, T., Klein, H.: Chemometric evaluation of concentration profiles from compounds relevant in beer ageing. Eur. Food Res. Technol. **215**(3), 235–239 (2002). https://doi.org/10.1007/s00217-002-0539-5
43. Wang, J., Li, Q.X.: Chapter 3—chemical composition, characterization, and differentiation of honey botanical and geographical origins. Volume 62 of Advances in Food and Nutrition Research, pp. 89–137. Academic Press (2011). https://doi.org/10.1016/B978-0-12-385989-1.00003-X
44. Wistuba, M., Rawat, A., Pedapati, T.: A survey on neural architecture search. CoRR, abs/1905.01392 (2019)

Citizen Data and Citizen Science: A Challenge for Official Statistics

Monica Pratesi

Abstract Citizen Data and Citizen Science are undoubtedly a challenge and an opportunity for Official Statistics. The paper follows the evolution in the production of statistics and indicators and gives some indications for the use of Citizen data in the production of indicators for monitoring of SDGs achievements.

Keywords Citizen data · Citizen science · SDGs

1 Introduction: Next Generation Data

In the last ten years official statisticians have been discussing the impact of big data and of new data sources in the production of Official Statistics (OS), highlighting many advantages and also disadvantages of their use. The main question was and is: "What is the future of Official Statistics in the big data era?".

A lot has been done for the use of big data in OS by the International and National Statistical Institutes (NSIs), including the Italian Statistical Institute (Istat). My contribution to the debate was initially on model-based estimates using big data sources [3, 4],then, in my capacity of President of the Italian Statistical Society (SIS), I intervened on the error profile of big data [7, 8]. Since last year, I have been focusing on Citizen Science, as the global process of digitization is so pervasive that times are mature for studying how to using and reusing Citizen Data in the production of OS [10].

As a matter of fact Official Statistics have always been evolving and the term "Trusted Smart Statistics" (TSS) was put forward by Eurostat and officially adopted by the European Statistical System (ESS) in 2018 to signify this evolution (Bucharest memorandum). In the debate many complex questions were posed: Could the use of big data, smart statistics, citizen data and citizen science in producing OS be a danger? Would OS be under attack either by discussions on trust or by competition

M. Pratesi (✉)
Department of Economics and Management, University of Pisa, Pisa, Italy
e-mail: monica.pratesi@unipi.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_12

with statistics produced with lower quality? Do official statisticians of the future need to be more than just data engineers [12]?

In this contribution I address the above questions, considering new data as the Next Generation Data, and schematizing the evolution track in data production process followed by NSIs. The evolution track—presented in Sect. 2 —has always guaranteed trust in data collection and processing. In Sect. 3 the challenges of Citizen Data (CD) and Citizen Science (CS) to OS are discussed. Finally, in Sect. 4, a project on the use of Citizen Science and Citizen Data to estimate BES indicators (Equitable Sustainable indicators), which will be implemented by Istat, is recalled.

2 The Evolution Process for Producing Statistics and Indicators

The mission of OS has always been to provide a quantitative representation of the society, economy, and environment for purposes of public interest, for policy design and evaluation and as a basis for informing the public debate. The production of modern OS is based on a system of scientific methods, regulations, codes, practices, ethical principles, and institutional settings that were developed through the last two centuries at the national level in parallel to the developments of modern states [13].

Figure 1 illustrates the evolution mechanism of the production process of a general OS system (engine), with its data sources (fuels) and User's information needs (accelerators). We see immediately that statistics and indicators are influenced both by fuels and accelerators. The rise of new data sources can give new fuels for statistics and indicators, but it can also act as a multiplier as it provokes new data information needs, becoming accelerators that stimulate further needs to be satisfied. Moreover, statistical methods and rules, for example, needed to guarantee privacy and trust on statistics and indicators produced, need obviously to adapt to the characteristics of the various data sources.

It is evident that this scheme of the evolution of the OS production process covers the current situation, but it is also valid for all the various breakthrough periods of data collection and statistical production of given NSIs.

For example, in Italy, a sudden change in information needs occurred after World War II. The government, policy makers and stakeholders needed new statistics to reconstruct the economic environment (Marshall and Fanfani Plans) and OS reacted by designing and carrying out surveys in different domains, in particular for the construction of National Accounts and developing new structured and standardized survey methods.

Therefore, the new scenario of data sources outlined in the introduction, moves the evolution mechanism, affecting for example the roles of the various stakeholders and their mutual relationships, also to address the questions of the National Recovery and Resilience Plan (NRRP). Summarizing, there is a need for timely reactions, both in terms of necessary reorganizations of the NSIs and their production and publications,

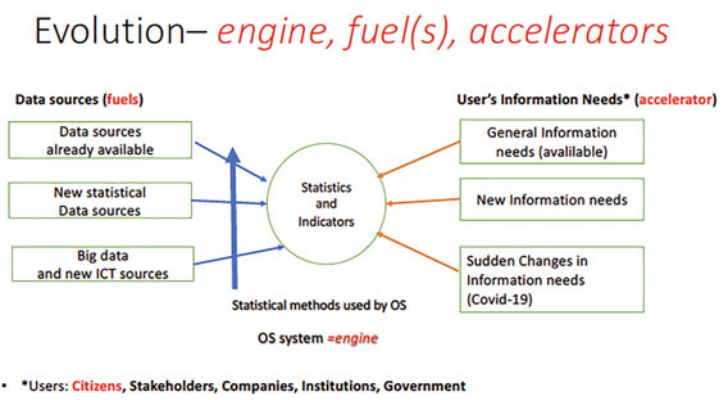


Fig. 1 Evolution of the production process of a general OS system

even with provisional results of the data collection process, as, for example, with Experimental Statistics.

3 Citizen Data and Citizen Science: A Challenge for OS

As already mentioned in the introduction, big data, smart statistics and citizens are inseparable: From smartphones, meters, fridges and cars to internet platforms, the data of most digital technologies is Citizen Data, that is the data of the citizens and on the citizens.

In addition to raising political and ethical issues related to privacy, confidentiality and data protection, the repurposing of big data calls for rethinking the relations between the citizens and the production of Official Statistics, if these are to be trusted. I am convinced that the future of Official Statistics does not depend only on the possibility to use new sources of data or new methods, but also on the possibilities that the new digital technologies offer to establish new relationships with the citizens. Their role is destined to evolve from that of respondents to that of collaborators and co-producers of Official Statistics data [14, 15].

There are two possible approaches. First, the possibility to exploit Citizen-generated data (CGD)—that are produced by non-state actors, particularly individual or civil society organizations—for official statistics purposes seems to represent a very promising avenue to collect timely and relevant new data. Privacy issues prevent citizens to fully disclose this kind of data, while their management and storage by privately owned digital platforms generate some remarkable concerns by citizens themselves on their correct protection. In order to fully exploit this kind of data, NSIs need to develop a better understanding on the way they are generated and how they can be made accessible for official purposes (Casarez-Crageda et al. 2020).

The second approach, that aims at the direct collaboration of citizens in the production of OS, following the principles of the Citizen Science (CS) implies the involvement of citizens along all the phases of the so-called data value chain: planning, collection, processing, analysis and use [5]. This is an important involvement that we can also link to the Post Normal Science approach [9].

The general opportunity for OS resides in gaining a new awareness of citizens in their participation to the process of official data production. Rethinking citizen involvement along the phases of the data value chain can help counter the trust deficit between citizens and governments and consequently establish a participatory data ecosystem [6].

In fact, as citizen data produced by digital devices (smartphones, meters, fridges and cars, internet platforms) are privately held, OS cannot directly access these data. OS can sign agreements with data owners (companies producing and selling the services distributed using smartphones, meters, fridges and cars, internet platforms) to have access to summarizations and tables from these data sources. In case of citizen generated data, OS can obtain the "consent forms" to use their personal data when using the digital device. Citizens act as data donors, hosting Apps released by NSIs on their devices (e.g. on mobile phones, meters, fridges) to collect data on mobility, energy consumption, expenditures.

The use phase in the data value chain requires an uptake stage that involves three activities: Connecting data to users; incentivizing users to incorporate data into the decision-making process; and influencing them to value data. The active involvement of citizens in the data production is a challenge for OS to reduce the gap between users and producers. In my opinion it would also have a positive feedback on statistical literacy, as the ability of data users to interpret and critically evaluate statistical information in a variety of contexts.

It is clear that the concept of citizen data and co-production raises practical and political questions that it is impossible to summarize here.

Moreover, CS produces data difficult to compare, the measures of precision are not clear. The challenge for OS resides in the rethinking the data collection process and on the concept of data quality. Traditional aspects inherent to the data production process and that are typical when NSIs conceive and govern it—such as accuracy, timeliness, representativeness, completeness, etc. should be rethought and enriched with the introduction of also other aspects. These last are important when the NSIs are not in the position to interfere in every aspect of the production process of the data as: Evaluation of self-selection bias, quality checks post-production, evaluation of potential use of the data. Issues as comparability across domains, coherence and benchmarking will be even more important than in traditional data production settings. Even if there are proposals to define the quality profile of citizen-generated data, there are not yet comprehensive and meaningful empirical studies.

To fill this gap we need to enhance research in OS to generate many Experimental Statistics, producing results also with unusual tools such as inference from nonprobability samples, data integration and data fusion of new and traditional data, model based and model assisted estimation methods.

This is true for all thematic areas where OS is called to produce data: From economic life, as consumption expenditure, earning and usage of disposable income to the aspects of daily life, like access to public services, life-long education, participation in social and cultural life, etc.

4 A Project on the Measure of the Quality of Citizen Data for the Compilation of SDGs Indicators

An important area, essential for regeneration and governance in this difficult moment marked by the pandemic, is that of the Sustainable Development Goals (SDGs).

The production of SDG indicators and their regular update is very demanding for NSIs that are struggling to balance financial constraints, the fast growing demand of new official statistics across different social, economic and environment domains with an increasing disaffection of respondents in reporting to NSI, despite their legal obligations. Since the production of SDG indicators may address data collection from specific target populations not always included in standard statistical business registers or consider information very difficult to collect based upon large scale official surveys or administrative data, CS initiatives and CGD clearly emerge in this area of statistical production as a unique and very promising solution.

SDGs are included in the government programs of European countries. A recent review highlighted how data collected through CS initiatives can feed an important part of indicators for monitoring Sustainable Development Goals [2]. Among the European countries where this practice is widespread, Italy is missing.

However, the possibility for NSI to successfully use CGD data for official statistical production in general, and for the set up and maintenance of SDG indicators in particular, has to meet the essential condition that their quality for statistical purposes can be carefully assessed and their potential biases corrected using a consistent methodological and statistical data processing approach.

Table 1 highlights some possible experimental settings that can be established by Istat to test the quality of CGD data for the compilation of SDG indicators.

The work is in progress and the settings in the table are the initial step of a complex experiment [11].

Istat has a leading role in Europe for the production of Equitable and Sustainable Well-being (BES) indicators, as witnessed by the recent presentation of the BES Report on March 10th, 2021. I believe that CS is a path to further improve the Institute's contribution.

Table 1 Experimental settings to test the quality of CGD for the compilation of SDG indicators

Area of reference of SDG indicators	Specific topic under investigation	Characteristics of units reporting CDG data	Availability of additional information	Methodology to test for the quality of CGD
Goal 1—Poverty	Poverty, material deprivation	Linkable to Business Register	Eu-Silc and Household Budget Survey	Record linkage, Statistical matching, Latent variables models
Goal 4—Education	Informal education (i.e. cinema, read books, theatre), soft skills	Linkable to Business Register	Labour Force Survey, educational registers and others administrative sources	Record linkage, Statistical matching, Latent variables models
Goal 2—Food security improve nutrition	Food waste	Linkable to Business Register	Household Budget Survey, Aspects of daily life	Record linkage, Statistical matching, Latent variables models

References

1. Cázarez-Grageda, K., Schmidt, J., Ranjan, R.: Reusing Citizen-generated data for official reporting a quality framework for national statistical office-civil society organisation engagement PARIS21 Working Paper (2020)
2. Fraisi, D., Campbell, J., See, L., When, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L., Masò, J., Penker, M., Fritz, S.: Mapping citizen science contributions to the UN sustainable development goals. Sustain. Sci. **15**, 1735–1751 (2020)
3. Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinizivillo, S., Pappalardo, L., Gabrielli, L.: Small area model-based estimation using big data sources. J. Off. Stat. **31**, 263–281 (2015)
4. Marchetti, S., Giusti C., Pratesi M.: The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy, in AStA Wirtschafts- und Sozialstatistisches Archiv: 57 **10**(2–3), 60−61 July 2016 (2016)
5. Nascimento, S., Iglesias, J.M.R., Owen, R., Schade, S., Shanley, L.: Citizen science for policy formulation and implementation, chapter 16 In: Hecher, S,. Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (2018), Citizen Science: innovation in Open Science, Society and Policy, UCI Press London (2018)
6. Misra, A. and Schmidt, J.: Enhancing trust in data – participatory data ecosystems for the post-COVID society, in Shaping The Covid-19 Recovery: Ideas From Oecd's Generation Y And Z © OECD 2020 (2020)
7. Pratesi, M.: Big data: the point of view of a Statistician, Etica e Economia, 12/4, (2017)
8. Pratesi, M.: Statistica: linguaggio sovradisciplinare per comprendere e dare valore ai dati talk in the Conference on "Data to Change" held on January 15, 2018 at the Italian House of Representatives, in Statistica & Società (2018)
9. Pratesi, M.: Parlare chiaro: statistica, dati e modelli, talk in "Parlare chiaro, i rischi della confusione dei numeri", online workshop, 30 aprile 2020, Università Politecnica delle Marche (2020)
10. Pratesi, M.: Official Statistics and Citizen Science, Seminar held March, 18, 2021, http://www.centrodagum.it/en/seminario-scuola-dei-dottorati-delle-scienze-sociali-universita-di-firenze/ (2021)

11. Pratesi, M., Ceccarelli, C., Menghinello, S.: Citizen generated data and Official Statistics: an application to SDGs indicators, Discussion paper n 274. Univ. Pisa Dep. Econ. Manag (2021)
12. Radermacher W. (2019), Governing-by-the-numbers/Statistical governance: Reflections on the future of Official Statistics in a digital and globalized society, Statistical Journal of the IAOS,
13. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M.: Trusted smart statistics: Motivations and principles. Stat. J. IAOS **35**, 589–603 (2019)
14. Ruppert, E., Grommé, F., Ustek-Spilda, F., Cakici, B.: Citizen data and trust in official statistics. Economie et Statistique/Economics and Statistics, No **505–506**, 179–193 (2018)
15. Ruppert, E.: Different data futures: An experiment in citizen data. Stat. J. IAOS **35**, 633–641 (2019)

Detecting States of Distress in Financial Markets: The Case of the Italian Sovereign Debt

Maria Flora and Roberto Renò

Abstract We evaluate a new test for distress, based on drift bursts, using the secondary market of Italian Treasury bonds in 2018. In May 2018, selling pressure in the secondary market due to a change of the Italian political scenario was not absorbed properly and caused overreaction, with direct costs in terms of harshened liquidity conditions, increased volatility and arbitrary wealth redistribution in favor of primary dealers. The new test proves to be a reliable tool for monitoring financial markets.

Keywords Drift burst test · Financial distress · Italian bonds

1 Introduction

Inefficient markets are a threat to investors. Market inefficiency of flash-crash type is a new form of financial fragility (in the Allen and Gale [1] sense) which plagues economically central markets. The flash crash of May 6, 2010 in the US stock market, triggered by a huge selling trade in the E-mini futures market [12], has attracted the attention of traders, institutions and academics (see e.g., [21, 24, 25]). The event shed light on a market vulnerability which appears to affect financial markets quite often, and increasingly over time [14, 17, 18, 23]. The natural question is: what is the impact of flash crashes on market activity and social welfare? The transient nature of these events may lead to think they do not matter much. Bank of England [4] writes: "Flash episodes have not, as yet, had financial stability consequences." On the other end, financial stability is defined as the "ability to facilitate and enhance economic processes, manage risks, and absorb shocks" [27].

Our research contributes to this literature by focusing on a deep crash that happened in the Italian sovereign bond markets on May 29, 2018, when a shock due

M. Flora
CREST, ENSAE, Institut Polytechnique de Paris, Palaiseau, France
e-mail: maria.flora@ensae.fr

R. Renò (✉)
Università di Verona, Verona, Italy
e-mail: roberto.reno@univr.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_13

to macro news (change of government after political elections) was not absorbed properly. Even if this event was somewhat slower, it had many similarities with flash crashes: it was characterized by a rapid, large and transient price decline, a slow recovery to nearly initial levels, all accompanied with severe and increasing liquidity evaporation as the crash evolved. We show that this event had large consequences on the market. Direct costs came from auctions which took place exactly at the bottom of the crash. The crash was indeed particularly unfortunate for Italian taxpayers, since the Treasury was auctioning at 11:00 of May 30. We estimate the money lost by the Treasury because of the crash (which involved one CCT and two BTPs, for a total of more than 6 billion euros offered) was roughly 0.45 billion euros (see Flora and Renò [17] for details).

Commenting on the crash, Financial Times[1] pointed at extreme volatility caused by a deterioration of market liquidity. Using a formal, recently developed statistical test, we show instead that the crash was due to a "drift burst" [14], that is a large (downward, in this case) trend in prices localized in a short time interval. The distinction between a volatility move and a drift move is not immaterial. Indeed, large volatility is possible even in an efficient and perfectly liquid market [17]. Large drift is instead typically associated with flash crashes, that is with inefficiency and market frictions.

Economic theory offers several explanations for the presence of drift bursts in financial markets, typically associated with the way traders strategically interact in a frictional environment. Frictions considered by theoretical papers include transaction costs, asymmetric information, market panic, market disruptions, and market fragmentation. For example, Grossman and Miller [19] predict a large and localized price decline (and subsequent reversal to the initial price level) in the presence of selling pressure looking for immediacy. The price decline of their model is proportional to the trade size and inversely proportional to the liquidity of the market. We document indeed that the crash of Italian bonds was associated with evaporating market liquidity. Reuters[2] reported: "Several banks have stepped back from primary dealing roles, partly due to regulatory pressures". The simple mechanism of Grossman and Miller [19] can be exacerbated by several frictional additions to the model. For example, Brunnermeier and Pedersen [10] show that opportunistic buyers could follow the initial sell orders to push the price downward in an illiquid market; Huang and Wang [20] show that market monitoring costs can exacerbate selling; Colliard [15] shows that flash crashes can be exacerbated by the presence of traders with superior information on liquidity, instead of fundamentals; and Menkveld and Yueshen [25] show that cross-arbitrage may break during a severe liquidity shock, and point at fragmented markets (like the Italian secondary market for Treasury bonds) as a potential source of flash crashes.

We use the Christensen et al. [14] test to study the deep crash in the Italian sovereign bond markets of May 29, 2018, when a shock due to macro news was not

[1] "Italian bonds' extreme volatility exposes liquidity strains", Financial Times, June 1, 2018.

[2] "Dwindling bond liquidity means Italy shock may be just a warning tremor", Reuters, July 3, 2018.

absorbed properly. We show that this abnormal price movement was associated with a drift burst, and that the secondary bond market liquidity was affected for several weeks afterwards. The crash in the Italian sovereign market lasted for a few hours, and propagated to the following day. The inefficiency implied a strong redistribution effect, since the crash happened precisely during a large Treasury auction, taking place on May 30 and issuing more than 6 billion euros (face value) of medium and long-term debt. An iconic example of the size of the event is the CCT with ISIN IT0005331878. This bond is a floating rate note with an additional fixed semi-annual coupon of 0.55% and maturity September 2025. Standard arbitrage theory dictates that, without credit risk, the clean price of the CCT should be 100 plus the discounted value of the fixed coupons. The volatility of this instrument, again without credit risk, should be almost zero. On the opening of May 28, before the crash started, the CCT was trading at the clean price of 96.51, below 100 because of the credit risk of Italy. At the bottom of the crash, reached at 11:38 a.m. of May 30, the CCT was trading at 87.96, that is almost 9% below the opening price at the beginning of the week. At the opening of June 1st, the CCT was trading again at 94, and the V-shaped path was completed. This transient crash was particularly unfortunate for Italian taxpayers, since the Italian Treasury was auctioning this specific CCT precisely at 11:00 a.m. of May 30, just 8 min before the price nadir. Other two large bond issuances took place in the same auction. Using a regime-switching model, we estimate that, during the auction of May 30, about 0.45 billion euros were transferred from Italian taxpayers to the primary dealers. We also document that the crash was associated with increased volatility and deteriorated liquidity conditions, that persisted for several months afterwards.

This case study contributes to the current debate on flash crashes and their financial stability implications. Indeed, the collapse of the Italian sovereign bond prices shows that flash-crash type behavior (as identified by a drift burst, or by a V-shape, see additional work in Flora and Renò [17]) can plague even economically central and large markets, and that this market inefficiency can span time scales much longer than a few minutes. Using both futures and cash data, Bouveret et al. [7] also document how liquidity deteriorated on the Italian sovereign bond market on May 29, 2018, when primary dealers retrenched from quoting bonds on the MTS interdealer platform. Moreover, these events can have sizeable effects in terms of welfare distribution. Finally, they have important financial stability implications in terms of systemic risk. The Spanish 10-year note yield soared to 1.62% on May 29, 2018, and it was 1.45% on May 22 and 1.34% on June 5, a clear signature of contagion. On the contrary, the German 10-year note yield plummeted at 0.257% on May 29, while it was 0.560% on May 22 and 0.420% on June 4, a clear signature of the propagation of the shock to other markets via the mechanism of flight-to-quality.

2 Empirical Analysis

The objective of this specific analysis is twofold. The first relates to the fact that economic theory predicts that large drifts (and reversals) are impossible in an efficient and liquid market, where prices should be semi-martingales (and local drift can only be "small"). Thus, showing that the event was associated with drift explosion compels to associate the crash with poor liquidity conditions and market inefficiency. The second objective is to show that the proposed drift burst test could have been used in this case as an early warning signal for the liquidity distress in the market.

We show that the crash of May 29 was associated to a large drift, not to large volatility, using nonparametric statistics to show that this evidence is robust to model specification. The common practice to detect financial distress is to look at price dispersion measures, such as volatility [5, 13], jumps [11], or simply at large returns [8]. The drift component of the price process is generally overlooked, in view of its negligibility with respect to a diffusive component in a conventional setup of locally bounded coefficients.

To identify large drifts we use the recent test statistic of Christensen et al. [14], henceforth COR. This is a non-parametric test which uses $n + 1$ log-price observations $X_0, \ldots, X_n$ observed at times $t_0, \ldots, t_n$. The test can be formally expressed, at time-point t, as:

$$T_t^n = \sqrt{\frac{h_n}{K_2}} \frac{\hat{\mu}_t^n}{\hat{\sigma}_t^n}, \tag{1}$$

where, for $t \in [0, T]$,

$$\hat{\mu}_t^n = \frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t_{i-1} - t}{h_n}\right) \Delta_i^n X \tag{2}$$

is a localized estimator of the drift, in which h_n is a bandwidth parameter measuring the extent of the localization, and K is a suitable kernel, while $\hat{\sigma}_t^n$ is a localized, pre-averaged and HAC-corrected [3] estimator of the spot volatility.

From a statistical point of view, the statistic T_t^n is expressing the ratio, for the log-return observed in a window of approximate length h_n, between the part of the return due to the drift and the part of the return due to the volatility. Under the null of bounded drift (that is, negligible price trends), the statistic is close to a normal distribution. A large value of the test statistic would thus signal an abnormal trend, or a "drift burst" in the COR language. In a large sample of liquid futures data (including US Treasury bonds), COR show that large values of the test statistics are almost always associated with large trading volume, and short-term price reversals, which are typical of flash crashes. In a related paper, using data on the French market, Bellia et al. [6] show that large values of the test statistics are unambiguously associated with reversals and evaporating liquidity. Flora and Renò [17] generalize the t-statistic in a V-statistic, which is proposed to test for market inefficiency.

We implement the test statistic (1) on selected Italian sovereign bonds in 2018 and 2019. We use tick data for a subsample of Italian government securities traded on the MOT (*Mercato Obbligazionario Telematico*), the electronic Italian-regulated limit order book market for sovereign, bank and corporate bonds. MOT is a retail exchange characterized by many transactions with small volume. It is the main retail trading venue by volume for Italian government bonds (8.25% of all trades on platforms in 2018), even if its volume is much lower than that of the two wholesale platforms, MTS Cash and MTS BondVision (91.24% of all trades on platforms in 2018).[3] Despite its relatively thin volume, the high number of transactions in the MOT guarantees that absence of cross-market arbitrage and fair security pricing is broadly guaranteed within the bid-ask spreads [28].

The daily trading schedule on the MOT is divided in two segments: an opening auction, from 8:00 a.m. to 9:00 a.m., followed by a continuous trading phase, from 9:00 a.m. to 5:30 p.m. The opening price is determined during the opening auction phase. We only focus on the continuous trading session, and exclude opening auction activity from the analysis. We select a representative set of Treasury bonds among BTPs (fixed coupon), CCTs (floating + fixed coupon), and BTPi (inflation linked bonds), namely:

- a 10Y BTP with maturity 2023, and 9% coupon rate (BTPs pay semi-annual coupon), BTP-1nv23 9%,
- a 30Y BTP with maturity 2029 and 5.25% coupon rate, BTP-1nv29 5.25%,
- a 30Y BTP with maturity 2040 and 5% coupon rate, BTP-1st40 5%,
- an inflation-linked 30Y BTP with maturity on 2035 and 2.35% coupon rate, BTPi-15st35 2.35%,
- a 7Y CCT with maturity 2022, and with an Euribor-linked coupon rate, CCT-Eu Tv Eur6m+0.7% Dc22.

We consider the period from January 1, 2018 (except for the last three securities, whose data start on the first issue date, that is January 31, 2018, February 28, 2018 and May 2, 2018, respectively) to May 30, 2019. The data were provided by Borsa Italiana S.p.A., and are recorded with millisecond time-stamps. All transactions are at the clean price.

We clean high-frequency transactions according to the following procedure:

1. for each trading day, we discard observations three times larger than the daily price median;
2. at the day-level, we then implement the Brownlees and Gallo [9] filter to filter out outliers: we keep the jth observation if

$$|p_j - \bar{p}_j(k)| < 3\sigma_j(k) + \gamma\,, \tag{3}$$

where $\bar{p}_j(k)$ and $\sigma_j(k)$ denote the δ-trimmed sample mean and standard deviation, respectively, of a neighborhood of k observations around j, while γ is the so-

[3] See CONSOB, Bollettino Statistico n. 14, June 2019, available at http://www.consob.it/web/area-pubblica/bollettino-statistico.

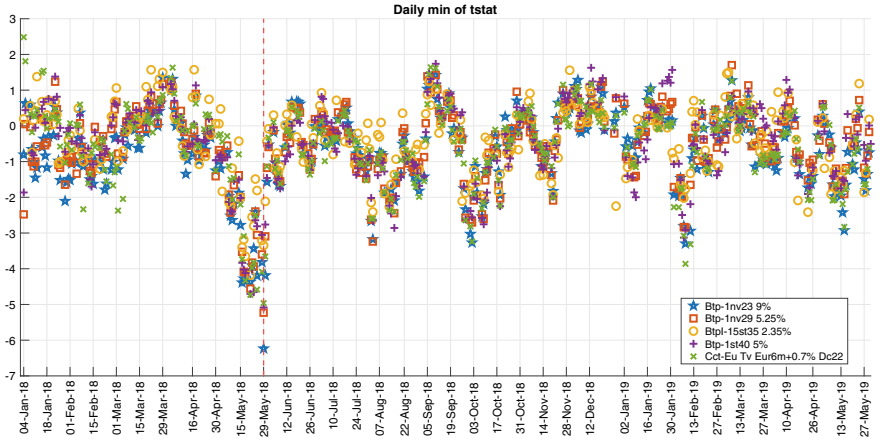


Fig. 1 Daily minimum of the drift burst test statistic proposed by Christensen et al. [14]. Large values of the test statistic signal market distress. The dashed-red line is May 29, 2018

called granularity parameter. We select $k = 50$ observations, $\gamma = 0.02$ (twice the minimum tick), and $\delta = 0.9$.

3. We aggregate transactions with the same time-stamp. We substitute simultaneous tick-by-tick prices with the volume-weighted average price, and simultaneous tick-by-tick volumes with the sum of the simultaneous volumes.

We choose a bandwidth h_n for the drift equal to 2 days, while we base the volatility on a 10-day bandwidth. We adopt a left-sided exponential kernel $K(x) = \exp(-|x|)$, for $x \leq 0$. Finally, for each trading day, we compute the minimum of the calculated t-statistics.[4]

Figure 1 reports the results of the drift burst test. The daily minimum of the test statistics T_t^n for the five instruments first crosses the -3 value approximately two weeks before the crash event. The five T_t^n all peak on May 29, with BTP-1nv23 9% being the most affected: the value of the t-stat is slightly below -6, and thus provides strong evidence of a dominating trend in prices. Few subsequent negative peaks in the test-statistics are observed in 2019.

What may cause a large value of the test-statistic? As mentioned, the neatest economic interpretation comes from the intermediation theory of Grossman and

[4] To deal with overnight gaps, we construct a new time vector for each time series, associated to the original one, where the time (in milliseconds) elapsed from the closing of day $t-1$ to the next available open price is equal to

$$\tilde{t} = \frac{\sigma_{overnight}}{\sigma_{intraday}} \hat{t}\,.$$

Here, $\sigma_{overnight}$ is the standard deviation of overnight returns, defined as the price appreciation or depreciation between market close of day $t-1$ and market open of day t, while $\sigma_{intraday}$ is the standard deviation of intraday returns, defined as the price appreciation or depreciation between market open and close of the same day. Finally, $\hat{t}$ is the time, in milliseconds, elapsed from market open (9 a.m.) and close (5:30 p.m.).

Miller [19]. In their model, a trader looking for immediacy is willing to sell to M market makers a volume s of a security. Market makers accept to trade immediately but with a price concession, which is given by the formula:

$$\frac{\mu}{\sigma} = \frac{s\gamma}{1+M}\sigma P_0, \tag{4}$$

where $\mu = E(P_1/P_0 - 1)$, P_0 is the initial price, P_1 is the price at which the market maker trades, σ is the standard deviation of the price move, and γ is the risk aversion of market makers. Intuitively, the price concession $P_1 - P_0$ will be larger when the traded size is larger, when volatility is larger, when risk-aversion is larger, and when the number of market makers (the unique measure of liquidity in the model) is smaller. Thus, large trading volumes generate the trends, overshooting and reversals which are typically observed when T_t^n is large. This setting remains valid when the initial selling is informed, as in this case (the signal coming from the change in the political scenario), since the selling pressure may still generate the transient effect predicted by the theory. Of course, alternative explanations are possible, as discussed in the introduction. Most of them are however just meant to reinforce this mechanism. About the connection of flash crashes to political uncertainty, Tsai [29] associates large changes with political turmoils, justified using the model of Kyle [22].

A key prediction of the Grossman and Miller [19] model is that the transient mispricing should be more severe in a market with poor liquidity. We compute two realized liquidity measures that can be inferred directly from transaction prices. The first is a measure of price impact which is close to the Amihud [2] measure (we modify it to take into account the irregular sampling of trades). The measure is implemented as follows. For each trade t and bond i in the sample, we compute

$$\text{Amihud}^*_{i,t} = \frac{|\Delta \log(p_{i,t})|}{V_{i,t}\sqrt{T_{i,t}}}, \tag{5}$$

where $\Delta \log(p_{i,t}) = \log(p_{i,t}) - \log(p_{i,t-1})$ is the log-return between trade $t-1$ and trade t, $V_{i,t}$ is the volume of the $t-th$ trade, and $T_{i,t}$ is the time, in milliseconds, between trade $t-1$ and trade t. Figure 2 shows the daily median of the measure in (5) for four bonds. As expected, the measure peaks in the crash week. Most importantly, the impact of the crash is to increase persistently the illiquidity measure in the market, with a transient effect that lasts several weeks after the crash.

The second statistic we employ is the one proposed in Roll [26], which we compute day-by-day:

$$\text{Roll} = 2\sqrt{\left(-\text{Cov}\left(\Delta p_{i,t}, \Delta p_{i,t-1}\right)\right)^+}, \tag{6}$$

where $\Delta p_{i,t}$ is the price difference between two consecutive transactions. This measure can be regarded as a proxy for the effective bid-ask spread: the higher its value, the higher the costs in terms of immediacy for the investors. Figure 3 shows the daily value of the Roll measure, which has a very similar dynamics to that of the Amihud* measure. Again, there is a marked spike during the crash, and a persistent impact on

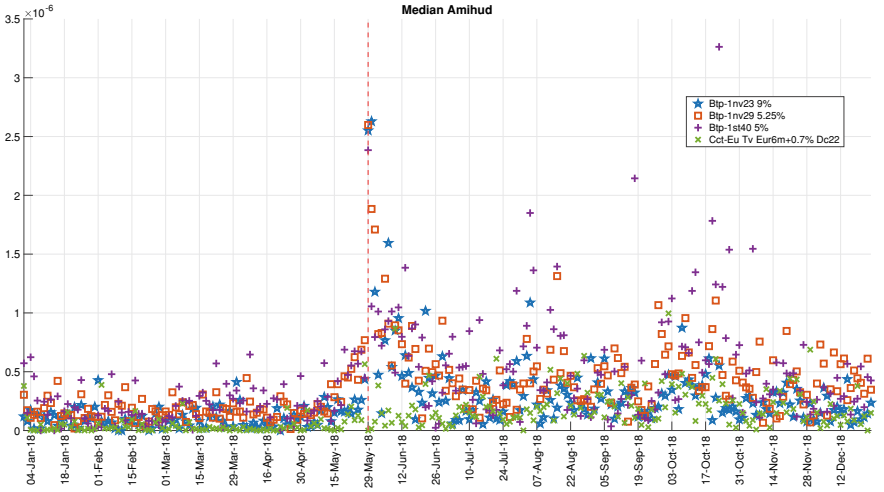


Fig. 2 Daily median of the Amihud* measure, as defined in Eq. (5). The dashed-red line is May 29, 2018

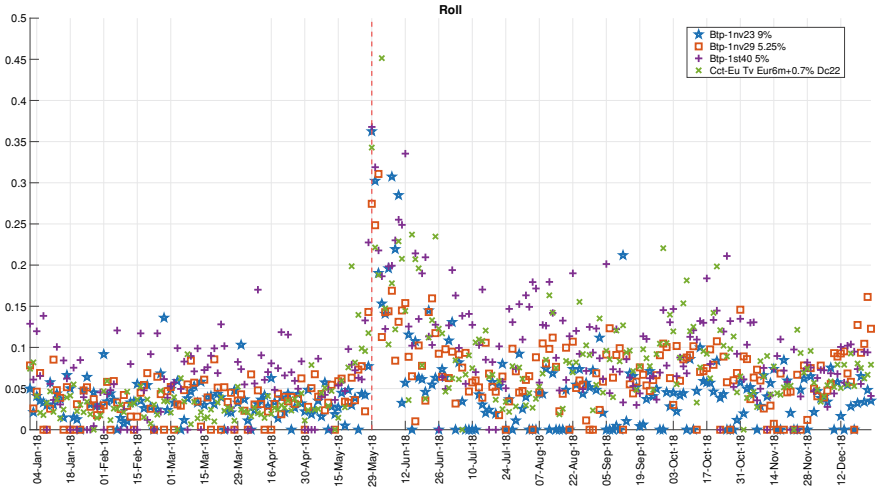


Fig. 3 Daily measures of the Roll illiquidity estimator. The dashed-red line is May 29, 2018

the market illiquidity, with a transient impact that lasts several weeks. Poor market liquidity is particularly relevant during price swings because of what Brunnermeier and Pedersen [10] call "predatory trading". According to this theory, an informed agent which observes a transient price decline in an illiquid market could exploit the larger price impact (see Fig. 2) to sell during decline, even if the fundamental price is higher. This is rational since the high price impact in a deprived market could make the price decline even more, making the later buying more convenient.

Importantly, the drift burst test could have been used to identify market distress from market transactions themselves. Using the precautionary threshold of -4.5, for the BTP 1Nov23 this line would have been crossed, using the same procedure described here in real time, at 9:53 of May 29, more than one day ahead of the auction of May 30, and never more in our sample. For the BTP 1Nv29, the line was crossed twice in the sample: at 17:23 of May 21, 2018 and at 10:26 of May 29. For the BPT 15St40, the line was crossed three times: 17:15 of May 21, 11:39 of May 23 and 10:36 of May 29. The CCT also crossed the line three times: 16:44 of May 21, 15:57 of May 25 and 10:31 of May 29. The BTPI never crossed the line, being the instrument with by far less transactions in our sample. Thus, a simple monitoring of the market would have at least informed market regulators that the market was distressed in the morning of May 29, and even with some "tremors" in the previous days.

3 Conclusions

We test a new technology, put forward by Christensen et al. [14], to detect distress in financial markets. We use it on a severe crash which occurred in the secondary Italian debt market, with huge consequences for the primary market and the Italian taxpayers, which Flora and Renò [17] quantify in a loss for the Treasury of around half a billon euros. Similar losses have been experienced during the COVID-19 pandemic [16] and can be associated to drift bursts as well. This finding is particularly important for financial stability, since it illustrates that the occurrence of phenomena similar to flash crashes is likely even in a systemic, allegedly liquid market like that for Italian Treasury bonds, and that their impact can be destructive. For example, the European Securities and Market Authority (ESMA) recently scrutinized the same event we analyze using data on the futures market [7]. Our research thus contributes to the debate of whether regulators should worry about the occurrence of flash crashes, and the conclusion of this paper is that they definitively should.

References

1. Allen, F., Gale, D.: Financial fragility, liquidity, and asset prices. J. Eur. Econ. Assoc. **2**(6), 1015–1048 (2004)
2. Amihud, Y.: Illiquidity and stock returns: cross-section and time-series effects. J. Finan. Markets **5**(1), 31–56 (2002)
3. Andrews, D.W.K.: Heteroscedasticity and autocorrelation consistent covariance matrix estimation. Econometrica **59**(3), 817–858 (1991)
4. Bank of England: Financial Stability Report (2019)
5. Bates, D.S.: How crashes develop: intradaily volatility and crash evolution. J. Finan. **74**(1), 193–238 (2019)
6. Bellia, M., Christensen, K., Kolokolov, A., Pelizzon, L., Renò, R.: High-frequency trading during flash crashes: walk of fame or hall of shame? Working paper (2019)

7. Bouveret, A., Haferkorn, M., Marseglia, G., Panzarino, O.: Flash crashes on sovereign bond markets—EU evidence. ESMA working paper, forthcoming (2021)
8. Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., Sokolov, K.: High frequency trading and extreme price movements. J. Finan. Econ. **128**(2), 253–265 (2018)
9. Brownlees, C.T., Gallo, G.M.: Financial econometric analysis at ultra-high frequency: data handling concerns. Comput. Stat. Data Anal. **51**(4), 2232–2245 (2006)
10. Brunnermeier, M., Pedersen, L.: Predatory trading. J. Finan. **60**(4), 1825–1863 (2005)
11. Calcagnile, L.M., Bormetti, G., Treccani, M., Marmi, S., Lillo, F.: Collective synchronization and high frequency systemic instabilities in financial markets. Quant. Finan. **18**(2), 237–247 (2018)
12. CFTC and SEC: Findings regarding the market events of May 6, 2010 (2010)
13. Christensen, K., Oomen, R.C., Podolskij, M.: Fact or friction: jumps at ultra high frequency. J. Finan. Econ. **114**(3), 576–599 (2014)
14. Christensen, K., Oomen, R.C.A., Renò, R.: The drift burst hypothesis. J. Econometr. **227**(2), 461–497 (2022)
15. Colliard, J.-E.: Catching falling knives: speculating on liquidity shocks. Manage. Sci. **63**(8), 2573–2591 (2017)
16. Ferrara, G., Flora, M., Renò, R.: The COVID-19 auction premium. Working paper (2021)
17. Flora, M., Renò, R.: V-shapes. Working paper (2020)
18. Golub, A., Keane, J., Poon, S.-H.: High frequency trading and mini flash crashes. Working paper (2017)
19. Grossman, S., Miller, M.: Liquidity and market structure. J. Finan. **43**(3), 617–633 (1988)
20. Huang, J., Wang, J.: Liquidity and market crashes. Rev. Finan. Stud. **22**(7), 2607 (2009)
21. Kirilenko, A., Kyle, A.S., Samadi, M., Tuzun, T.: The flash crash: high frequency trading in an electronic market. J. Finan. **3**, 967–998 (2017)
22. Kyle, P.: Continuous auctions and insider trading. Econometrica **43**, 1315–1335 (1985)
23. Laly, F., Petitjean, M.: Mini flash crashes: review, taxonomy and policy responses. Bull. Econ. Res. **72**(3), 251–271 (2020)
24. Madhavan, A.N.: Exchange-traded funds, market structure and the flash crash. Finan. Anal. J. **68**(4), 20–35 (2012)
25. Menkveld, A.J., Yueshen, B.Z.: The flash crash: a cautionary tale about highly fragmented markets. Manage. Sci. **10**(10), 4470–4488 (2019)
26. Roll, R.: A simple measure of the implicit bid-ask spread in an efficient market. J. Finan. **39**, 1127–1139 (1984)
27. Schinasi G.J.: Defining financial stability. IMF working paper (2004)
28. Schneider, M., Lillo, F.: Cross-impact and no-dynamic-arbitrage. Quant. Finan. **19**(1), 137–154 (2019)
29. Tsai, I.-C.: Flash crash and policy uncertainty. J. Int. Finan. Markets Inst. Money **57**, 248–260 (2018)

Forecasting Combination of Hierarchical Time Series: A Novel Method with an Application to CoVid-19

Livio Fenga

Abstract Multiple, hierarchically organized time series are routinely submitted to the forecaster upon request to provide estimates of their future values, regardless the level occupied in the hierarchy. In this paper, a novel method for the prediction of hierarchically structured time series will be presented. The idea is to enhance the quality of the predictions obtained using a technique of the type forecast reconciliation, by applying this procedure to a set of optimally combined predictions, generated by different statistical models. The goodness of the proposed method will be evaluated using the official time series related to the number of people tested positive to the SARS-CoV-2 in each of the Italian regions, between February 24th 2020 and August 31th 2020.

Keywords ARIMA model · ARFIMA model · Exponential smoothing model · Forecast reconciliation · Forecast combination · Model uncertainty · SARS-CoV-2 · Theta method

1 Introduction

In many applications, it is often the case that accurate forecasts are needed for time series showing an inherent hierarchical structure. For example, in economics the forecaster is routinely asked to provide separate forecasts for the industrial production index at the most aggregated level as well as for specific (sub–) classes of economic activities. The estimation of the future demand of domestic tourism usually follows a geographical proximity criterion, based on which the related time series are organized (and predicted) according to homogeneous groups. Sometimes, emergency situations

The original version of this chapter was revised: The incorrect coauthor names have been removed throughout the chapter. The publisher's correction to this chapter is available at https://doi.org/10.1007/978-3-031-16609-9_33.

L. Fenga (✉)
University of Exeter, Stocker Rd, EX4 4PY Exeter, UK
e-mail: l.fenga@exeter.ac.uk

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_14

require close monitoring of the spread of a disease not only at a national but also at a regional level, e.g. in order to set up more appropriate countermeasures for elderly and chronically ill people. These are all cases where a single line of hierarchy generates the overall structure of the data which therefore is referred to as "hierarchical time series". The present paper is concerned with the forecast of such data structures. While on one hand it is always possible to disregard the underlying hierarchical arrangement and thus carry out the prediction exercise considering each time series singularly, on the other hand, by doing so, it is very unlikely for the resulting higher level forecasts to be equal to the sum of the lower level ones. It goes by itself that such a situation is not acceptable in many instances, e.g. in the field of official statistics where aggregation consistency is generally a *conditio sine qua non* and even a slight misalignment needs to be dealt with. Therefore, many combination techniques have been designed to preserve the needed adding-up conditions, by accounting for the position occupied in the hierarchy by each and every time series, regardless their level of (dis–)aggregation. However, this type of approach—usually referred to as "forecast reconciliation"—is in general dependent on the statistical model *a priori* chosen to carry out the forecasting exercise. Undoubtedly, this choice might negatively impact the quality of the generated forecasts, e.g. by conveying not negligible amount of uncertainty into the analysis. This is especially true in the case of real-life data—where problems, such as small sample size, noise and systematic and/or non systematic errors—might hinder the choice of the "right" statistical model.

Motivated by this, in the present paper a method built upon the forecast reconciliation procedure devised by [21] will be presented. In other words, a joint hierarchical forecasting system will be formulated, where an additional optimality condition, derived in a multi-model setup of the type forecast combination, drives the choice of the "best" statistical model generating the predicted values. The final part of the procedure is designed to lower the bias of the selected forecasts. The main novelty of the method is that the forecast combination is applied directly on forecasts which have already been reconciled. In essence, it is an optimization procedure articulated in four steps: one performed at a cross-section level (reconciliation), two at a cross-model level (forecast combination) and the final one on the chosen prediction vector, for bias adjustment purposes. Surprisingly, to the best of the author's knowledge, this is the first attempt of this sort in the case of cross-sectional hierarchical time series.

The rest of the paper is structured as follows: Sect. 2 is devoted to the literature review concerning the two statistical methods the proposed procedure is based on, which will be detailed in the following Sect. 3. The proposed method, as well as its justification, will be respectively illustrated in Sects. 4 and 5. The following Sect. 6 will be devoted to an extensive empirical application, carried out using the official Italian data related to the SARS-CoV-2 positive cases, which will demonstrate the validity of the proposed approach. Section 7, containing the conclusions and the future directions of this work, will end the paper.

2 Literature Review

As already mentioned, the proposed procedure is based on two classes of methods, usually referred to as "forecast reconciliation" and "forecast combination". The former serves the purpose of achieving aggregation consistency of individual, aggregation inconsistent, forecasts whereas the latter will be employed to combine different reconciled forecasts, each of them generated according to different statistical models.

A rigorous and theoretically sound investigation on forecasts combination dates back to the late 60s—with the famous seminal paper by [6]. Here, the Authors showed that the combination of forecasts often leads to a better forecast accuracy and, by doing so, provided an alternative way to the notion that a "best" method exists and can be identified. Ever since this paper, the integration of a number of forecasts, independently estimated on a single time series, has attracted a great deal of research interest and, as a result, a vast literature is today available. Much of it is aimed at presenting empirical applications documenting the appealing features of this approach, which in many cases can improve even upon the best individual forecast, in terms of forecast risk, forecast error variance and consistency between in-sample and out-of-sample error distributions, as pointed out by [5]. This can happen for a variety of reasons, many of them related to the fact that the choice of the "right" model, in general, implies the injection of not negligible amounts of uncertainty into the analysis [9, 10]. In the same line of thinking, many Authors, see for example [25, 35, 36], emphasize the dangers related to misspecification errors, which, on the other hand, can be mitigated by combining the forecasts yielded by a number of models. In support of this argument, there are a number of studies which show that it is very unlikely, using a well calibrated portfolio of models, that one of them consistently dominates the others across the whole prediction window. Such an argument is consistent with the view that the "true" underlying data generating process is, saved for trivial or lab controlled cases, way too complicated to be adequately captured by a single model. This is the position, for example, of [8], according to whom the data can never support, and we can never identify, the "true" model. Therefore, the selection of a statistical model is more realistically the process of identifying the best approximating one. Once defined models as approximations, the concept of the identification of the "true" one ceases to be decisive in favor of approaches pursuing, in the first place, the goal of achieving good forecasting performances.

Many practical uses of forecast combination are discussed in the excellent work of [11], where the author covers a wide spectrum of applications, ranging from economics, demography and politics to meteorology and outcomes of football games. In the same spirit is the more recent paper by [26], which presents a classification of 174 articles focusing on forecast combination. In particular, new applications are reported from different sectors, such as commercial, tourism, urban traffic, betting market and propagation of successful innovations. The analysis of the outcomes of the M3 forecast competition, discussed in [28], goes in favor of the forecast combination

approach, which on average proved to outperform the methods individually applied. Same conclusion applies to the more recent M4 forecast competition, discussed in [27], where 12 out of a group of 17 most accurate methods are based on the combination of forecasts. Other considerations in favor of this approach are more closely related to the features of the time series under investigation. For instance, in many real-life cases they are affected by structural breaks, induced by a variety of factors whose real-time detection is generally difficult to achieve. However, in a multi–model setup it is not unreasonable to have models showing different degrees of ability in handling such events. Such a situation can translate into gains in terms of forecast accuracy, as it has been argued not only since the very beginning, in the above mentioned paper by Bates and Granger, but also in more recent times, by, among others, [12, 25, 32]. All these Authors concur on the premise that, on average, combining the forecasts yielded by models with different reaction times to a given intervention—and thus requiring stretches of post-break data of different length—can do a better job than individual models. For what said, it comes at no surprise that such appealing results might lead to a change in the perspective many researchers and practitioners look at the forecasting methods, i.e. from model selection—based on the assumption of the existence of one, "true" data generating process—to model averaging.

When the data set under investigation show a hierarchical structure, forecast combination techniques (but this holds true for any univariate forecasting procedure) are insensitive to the level of aggregation at which they are applied. Consistently, in such cases, the independent forecasting of the component time series is always possible, even if not advisable, due to the very likely lack of consistency occurring between the sum of the predictions generated at one level with those available at the level above. But this is not the whole story: by applying forecasting procedures at the components level, rather than limit them to the most aggregate one, it is possible to adequately capture the data covariance structure and thus achieve not negligible gains in terms of quality of the predictions. This fact has been pointed out, *inter alia*, by [14, 23, 29]. Their conclusions are related to two of the most traditional approaches, usually referred to as Top-Down and Bottom-Up (see, for example, [4, 24, 33]). While the former envisions a two-step procedure—i.e. the forecasting is first performed at the top level and then, by disaggregating these data based on the historical percentage of each data point, within the whole group—in the latter each and every time series is first individually predicted and then all the forecasts are summed up. There is another approach which has gained widespread acceptance over the years, known as "middle-out". It can be considered an extension of the top-down approach, since the forecast is first generated at two separate levels (upper and lower) and then combined in a proper manner to form a composite forecast. It is worth outlining how all of these methods can be considered sub-optimal insofar they neglect the correlation structure existing among the series belonging to the same level. Finally, a more recent approach known as optimal combination, envisions a two-step procedure where first the sequential and exhaustive forecast of each and every time series is independently performed and then—by optimally combining the predicted values obtained—are aggregated

to achieve consistency across the hierarchical levels (reconciliation). Theoretically, cross-level coherency can be attained by means of Generalized Least Squares (GLS) whose employment, however, turns out to be unfeasible due to the unidentifiability of the covariance matrix of the reconciliation errors [39]. However, other methods, e.g. of the type OLS [2] or WLS [22] can be used to circumvent this hurdle.

The method proposed in the present paper is of the type mixed, in the sense that exploits the approaches related to both forecasts reconciliation and forecasts combination. It is noted how mixed methods of this sort are not often encountered in literature. Such a situation might be due to the relatively recent introduction of reconciliation methods capable, unlike more traditional procedures, of accounting for the correlation structures among the series within a given hierarchical level, and thus able to deliver better performances. On the other hand, the need of a unified framework combining these two approaches has been recently brought up by [13], which discussed an *ad hoc*, bi-dimensional (cross-sectional and temporal) procedure built upon a recent proposal by [39]. Their method employs all the summation constraints arising in the cross-temporal hierarchical structure to reconciliate the base forecasts, using simple projections in a suitable linear space. On the other hand, the recent proposal by [34] envisions a common framework where both forecast reconciliation and combination of forecasts generated by multiple models work together. Their method arises from the consideration that base forecasts should not be derived from a single method but a combination of methods. In the same direction goes the early work by [38], where both combination and reconciliation of the forecasts are applied in a two-step procedure, i.e. "first one comes up with the best possible forecasts for the time series without worrying about aggregation consistency and then a reconciliation procedure is used to make the forecasts aggregate consistent". As it will be seen, the procedure illustrated in the present paper significantly differs from the above mentioned ones, in that different base forecasts are generated according to an arbitrary, pre-specified portfolio of statistical models, so that the combination exercise is performed on a set of already reconciled forecasts. Finally, while these authors focus on temporal aggregation this paper considers cross-sectional aggregations.

3 The Framework

In this Section, an explanation of the framework within which the proposed method operates is given. In particular, the hierarchy structure of reference along with the employed reconciliation method are illustrated. The forecast combination part will be explained using two real-life examples of portfolios—which will be both used in the empirical Section—related to the statistical prediction models and the forecast combination techniques entertained.

3.1 Hierarchical Cross-Sectional Reconciliation: The Chosen Method

This paper focuses on structures of the type summation constrained, in the sense that the underlying hierarchical structure of a given $m-$dimensional time series $\boldsymbol{x}_t$, arises by summing up the bottom-level series into the higher ones. Figure 1 is an example of such a structure, under the condition that the constraints $x_t = x_{a,t} + x_{b,t}$, $x_{a,t} = x_{aa,t} + x_{ab,t} + x_{ac,t}$ and $x_{b,t} = x_{ba,t} + x_{bb,t}$ are all satisfied.

Formally, we have that the observed data $\boldsymbol{x}_t$—as well as their estimated future values, defined as $\boldsymbol{x}_h$; $h = 1, 2, \ldots, H$, with H the prediction horizon—lie in the summation-coherent subspace $\{\mathcal{U}\}$; $\forall t = 1, 2, \ldots, T$ and $\forall h = 1, 2, \ldots, H$. The prediction step subscript h has been omitted in Fig. 1, for the sake of a better readability. In total, this hierarchy contains $m = 8$ time series, $n = 5$ of which are the lowest level time series, which therefore constitute the highest level of disaggregation. The observed series $\boldsymbol{x}_t \in \mathbb{R}^m$ can be broken down as follows: $\boldsymbol{x}_t = [\boldsymbol{u}_t', \boldsymbol{b}_t']'$, where $\boldsymbol{b}_t' \in \mathbb{R}^n$ and $\boldsymbol{u}_t' \in \mathbb{R}^{m-n}$ respectively contain the data pertaining to the bottom and upper series. Therefore, according this representation, the structure of Fig. 1 (omitting the subscript t) can be broken down as follows: $[\boldsymbol{u}_t', \boldsymbol{b}_t']' \equiv [x, x_a, x_b, x_{aa}, x_{ab}, x_{ac}, x_{ba}, x_{bb}]'$, $\boldsymbol{u}_t' \equiv [x_a, x_b]'$ and $\boldsymbol{b}_t' \equiv [x_{aa}, x_{ab}, x_{ac}, x_{ba}, x_{bb}]'$. The hierarchical structure—satisfying $\boldsymbol{x} \subset \{\mathcal{U}\}$—is induced by the summing matrix $\boldsymbol{\mathcal{S}}$ of dimension $m \times n$ such that $\boldsymbol{x}_t = \boldsymbol{\mathcal{S}}\boldsymbol{b}_t$. Formally: $\boldsymbol{x} \subset \{\mathcal{U}\} \iff \boldsymbol{x}_t = \boldsymbol{\mathcal{S}}\boldsymbol{b}_t$ (the symbol $\iff$ replacing the locution "if and only if"). The $\boldsymbol{\mathcal{S}}$ matrix for the hierarchy in Fig. 1 is as follows:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the symbols $\tilde{}$ and $\hat{}$ respectively to refer to the case of coherent and base (generally non coherent) forecasts, the reconciliated forecast $h-$ step ahead can be expressed as proposed by [21], i.e.

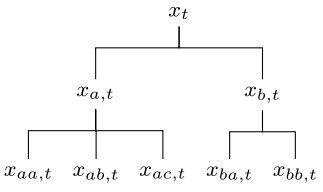


Fig. 1 A two-level hierarchical structure

$$\tilde{\boldsymbol{x}}(h) = \boldsymbol{\mathcal{S}}\boldsymbol{P}\hat{\boldsymbol{x}}(h), \tag{1}$$

for some appropriately chosen matrix $\boldsymbol{P} \in \mathbb{R}^{m\times n}$. Assuming unbiased base forecasts, the best linear unbiased revised forecasts (minimizer of the sum of the variances of the lower hierarchical level) are given by Eq. 1 with

$$\boldsymbol{P} = (\boldsymbol{\mathcal{S}}'\boldsymbol{W}^{-1}\boldsymbol{\mathcal{S}})^{-1}\boldsymbol{\mathcal{S}}'\boldsymbol{W}^{-1} \tag{2}$$

and thus (see [37], Theorem 1)

$$\tilde{\boldsymbol{x}}(h) = \boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}'\boldsymbol{W}^{-1}\boldsymbol{\mathcal{S}})^{-1}\boldsymbol{\mathcal{S}}'\boldsymbol{W}^{-1}\hat{\boldsymbol{x}}(h), \tag{3}$$

where $\boldsymbol{\mathcal{S}}$ is as above defined and $\hat{\boldsymbol{x}}(h)$ and $\tilde{\boldsymbol{x}}(h)$; $h = 1, 2, \ldots, H$ represent respectively the set of H predictions independently generated and the ones made coherent. Finally, $\boldsymbol{W}$ is the positive definite covariance matrix of the base forecast errors, i.e. $\hat{\boldsymbol{e}}_t(h) = \hat{\boldsymbol{x}}_t(h) - \boldsymbol{x}_t(h)$, so that $\boldsymbol{W}(h) = \mathbb{E}[\hat{\boldsymbol{e}}_t(h) - \hat{\boldsymbol{e}'}_t(h)]$. As shown by [39], matrix $\boldsymbol{W}(h)$ appears in the equation for the estimation of the error variance of the reconciled forecasts, i.e.

$$\boldsymbol{V}(h) = Var[\boldsymbol{x}(T+h) - \tilde{\boldsymbol{x}}(h)] = \boldsymbol{\mathcal{S}}\boldsymbol{P}\boldsymbol{W}(h)\boldsymbol{P}'\boldsymbol{\mathcal{S}}', \tag{4}$$

whose diagonal elements are the variances of the forecast errors. Their minimization can thus be performed in terms of the trace of $\boldsymbol{V}(h)$ and given by Eq. 2 (therefore, this method is called Minimum Trace Estimator). Unfortunately, as proved by the same Authors, $\boldsymbol{W}$ is not identifiable, therefore, in the empirical section, the work around proposed by them will be adopted. In essence, it is assumed $\boldsymbol{W}_h = k(h)\text{diag}\hat{\boldsymbol{W}}_1$; $\forall h$ and assuming $k(h) > 0$ and denoting with $\boldsymbol{W}_1$ the forecast errors covariance matrix estimated at horizon $h = 1$—i.e. $\hat{\boldsymbol{W}}_1 = \frac{1}{T}\sum_1^T \hat{\boldsymbol{e}}_t\hat{\boldsymbol{e}}_t{}'$—and with K is an unknown constant depending on the time horizon h.

3.2 The Forecast Combination Methods Adopted

As already mentioned, the proposed method uses a set of combination methods, out of which the winner is selected according to a suitable loss function. In many empirical studies, it is shown how forecast combinations on average delivers better performances than methods based on a single forecasting statistical models. The theoretical validity of this approach is rooted in the assumption that the dimension of the sample sizes available in real-life applications are usually finite and, as a result, the correct specification of the "true" underlying data generation process is not attainable.

In what follows it is assumed $\boldsymbol{x}_t$ to be the variable of interest and that $\boldsymbol{f}_t = (f_{1t}, f_{2t}, \ldots, f_{Nt})'$ are the N, not perfectly collinear, available forecasts, whose

combination is expressed as $f = \sum_{i=1}^{N} w_i f_{it}$, or, equivalently, $f = \boldsymbol{f}_t' \boldsymbol{w}$, being w's the combination weights.

The first method considered in this paper is of the type simple average. Despite its inherent simplicity (it ignores the correlation structure of the forecast errors) this method has been adopted given its ability, proved true in many cases, to "dominate more refined combination schemes aimed at estimating the theoretically optimal combination weights" [3]. The simple average assigns equal weights to all predictors, i.e. $\boldsymbol{w}^{sa} = \frac{1}{N}$ and thus the combined forecast is

$$f = \boldsymbol{f}_t' \boldsymbol{w}^{sa}.$$

In the second method chosen, the forecast combination weights:

$$\boldsymbol{w}^{ols} = (w_1, w_2, \ldots, w_N), \tag{5}$$

along with the intercept b, are computed using ordinary least squares (OLS) regression [18], i.e.

$$f = b + \boldsymbol{f}_t' \boldsymbol{w}^{ols}. \tag{6}$$

The third method applied—of the type Least Absolute Deviation (LAD)—is a modification of the OLS method, and it is expressed as in Eq. 6, replacing the superscript ols with lad. Since the method of least squares assigns heavy weights on the error terms, the more robust estimator LAD—of the type Gauss–Laplace (see, e.g. [16])—minimizes the absolute values and not the squared values of the error term. This features is particularly useful when the error term is generated by distributions having a infinite variance (fat tails) caused by outliers in the disturbance term.

Finally, a modification of the method proposed by [30], built upon an earlier methodology of [6], is our fourth approach. Let $\sum$ be the positive definite matrix of the mean squared prediction errors (MSPE) of $\boldsymbol{f}_t$ and $\boldsymbol{g}$ is an $N \times 1$ vector of $(1, 1, \ldots, 1)'$ their method relies on a constrained minimization of the MSPE under the normalizing condition $\boldsymbol{g}'\boldsymbol{w} = 1$. The resulting combination of weights is

$$\boldsymbol{w}^{ng} = \frac{\sum^{-1} \boldsymbol{g}}{\boldsymbol{g}' \sum^{-1} \boldsymbol{g}},$$

so that the combined forecast is

$$f = \boldsymbol{f}_t' \boldsymbol{w}^{ng}. \tag{7}$$

However, unlike the original method, the variant employed here follows the proposal by [20], which does not impose the prior restriction that the matrix $\sum$ is diagonal.

3.3 The Entertained Statistical Models

Before delving into the proposed method, a quick presentation of the forecasting methods employed in the empirical section is in order. The first two statistical models considered are of the type ARIMA (Auto Regressive Fractional Moving Average) [7] and $ARFIMA$ (Auto Regressive Fractional Moving Average) [17, 19]. Being the latter a generalization of the former, the two models will be presented conjointly.

$ARFIMA$ (Auto Regressive Integrated Fractional Moving Average) models are useful in circumstances where the underlying stochastic process exhibits hyperbolic decay patterns in their estimated autocorrelation function. ARFIMA-type processes are usually expressed as follows:

$$\mathbf{\Phi}(B)(1-B)^d x_t = \mathbf{\Theta}(B)\varepsilon_t; \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2),$$

where d is a parameter—assumed to take non-integer values in the difference operator $(1-B)^d$, with B identifying the backward operator, that is $B^k x_t = x_{t-k}$. The fractional differencing operator is defined by the binomial expansion $(1-B)^d = \sum_0^\infty \binom{d}{i} (-B)^i$. The process is stationary and invertible if the roots of the autoregressive polynomial of order p, $\mathbf{\Phi}(B) = 1 - \phi_1 B - \phi_2 B^2, \ldots, -\phi_p B^p$, and the order q moving-average part, $\mathbf{\Theta}(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$, lie outside the unit circle with $|d| < 0.5$.

ARFIMA models generalize the ARIMA(p,d,q) representation where the parameter d is constrained to integer values. This type of model has been designed to capture approximately parabolic decay patterns of the empirical autocorrelation function. As such, they are suitable to model persistence structures embedded in the underlying stochastic process of the type short-memory.

Theta method—the third forecasting model considered—is a powerful class of models which have been proposed by [1]. Define with the symbol ∇ the difference operator—i.e. $\nabla x_t = x_t - x_{t-1}$, x_t being the original time series—this method is the solution of the equation

$$\nabla^2 z_t(\theta) = \theta \nabla^2 x_t; \quad t = 3, 4, \ldots, n, \tag{8}$$

with $z_t(\theta)$'s analytical solution reading as following: $z_t(\theta)\theta x_t + (1-\theta)(A_n + B_n t)$; $t = 1, 2, \ldots, n$, where A_n and B_n are the minimum square coefficient of a linear regression equation of the series x_t against $\mathbf{1}_n$, i.e. the vector of ones of length n. These terms are given by $A_n = \frac{1}{N}\sum_{t=1}^n x_t - \frac{n+1}{2} B_n$ and $B_n = \frac{6}{n^2-1}\left[\frac{2}{N}\sum_{t=1}^N x_t t - \frac{1+n}{n}\sum_{t=1}^N x_t\right]$. Finally, the initial values z_1 and z_2 in Eq. 8 are estimated by minimization of $\sum_{t=1}^n (|x_t - z_t(\theta)|^2)$.

The fourth and last model employed are of the type exponential smoothing, proposed in 1944 by Robert G. Brown, a US Navy operations research analyst [15]. Specifically, two schemes have been employed here: Additive Holt Error Model (AEM) and multiplicative Holt Error Model (MEM) (as it will be seen, the procedure automatically will select the "best" one). As for the AEM, let

$\mu_t = \hat{x}_t = l_{t-1} + b_{t-1}$ be the one-step ahead forecast of the observed time series x_t generated by the forecasting equation $x_t = l_{t-1} + b_{t-1} + \varepsilon_t$, being l_t a measure of the level of the series, b_t an estimate of the slope (or growth) at time t and $\varepsilon_t = x_t - \mu_t$ the one-step-ahead forecast error, referred to the time t. The level and slope equations for AEM are respectively represented as

$$\begin{aligned} l_t &= l_{t-1} + b_{t-1} + \alpha\varepsilon_t \\ b_t &= b_{t-1} + \beta(l_t + l_{t-1} - b_{t-1}) = b_{t-1} + \alpha\beta\varepsilon_t. \end{aligned}$$

By re-expressing the error term ε_t as $\varepsilon_t = \frac{(x_t - \mu_t)}{\mu_t}$ (relative errors), the forecast, level and slope equations for the MAM model are as follows:

$$\begin{aligned} l_t &= (l_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t) \\ b_t &= b_{t-1} + \beta(lt - 1 + b_{t-1})\varepsilon_t. \end{aligned}$$

In the above two sets of equations α and β are the model parameters to be estimated.

4 The Proposed Method

Let us indicate with the symbols $\mathcal{R}$ and $|\cdot|$ respectively a suitable reconciliation method and the cardinality function (assuming the number of elements in a given set to be finite, $|\cdot|$ simply returns the number of the elements belonging to that set). Let the symbol **ncol** identify the function which, applied to a given matrix, returns its number of columns and $\boldsymbol{\mathcal{M}} \equiv \{\mu_1, \mu_2, \ldots, \mu_M\}$ and $\boldsymbol{\mathcal{D}} \equiv \{\delta_1, \delta_2, \ldots, \delta_D\}$ respectively the set of $|\boldsymbol{\mathcal{M}}| = M$ prediction models and the set $|\boldsymbol{\mathcal{D}}| = D$ of forecast combination methods entertained, both arbitrarily chosen. Once applied to the time series of interest $x_t;\ t = 1, 2, \ldots, T$, each model $\{\mu_j \in \boldsymbol{\mathcal{M}};\ j = 1, 2, \ldots M\}$ generates a set, called $\boldsymbol{\mathcal{F}}^H$, made up with M H–step ahead predictions, i.e.: $\{\boldsymbol{\mathcal{F}}^H(\mu_j);\ j = 1, 2, \ldots M\}$. Each of the elements of this set is a base forecasts, in the sense that it is generated by individually applying a given statistical model μ_j to the observed time series without any attempt of reconciliation.

Each of these M elements in $\boldsymbol{\mathcal{F}}$ (the M forecast vectors) is individually reconciled through the reconciliation procedure $\mathcal{R}$, i.e. $\left\{\mathcal{R}(\boldsymbol{\mathcal{F}}(\mu_j);\ j = 1, 2, \ldots M)\right\}$ (the superscript h is omitted for brevity). At this point, the resulting set $\left\{\boldsymbol{\mathcal{P}}(\mu_j);\ j = 1, 2, \ldots M\right\}$ of M model–dependent reconciled forecasts (first optimization) is optimally combined by applying each method in the set $\boldsymbol{\mathcal{D}}$ to any possible combination (without repetition) of order $\{k = 1, 2, \ldots, M\}$ to the set $\boldsymbol{\mathcal{P}}$ (second optimization). The resulting set $\boldsymbol{\mathcal{Z}}$—with cardinality $(|\boldsymbol{\mathcal{D}}| * \sum_{k=1}^{|\boldsymbol{\mathcal{P}}|} \binom{M}{k})$—contains all the possible combinations—$\forall k$-order—of the model-dependent reconciled forecasts. The third optimization step is carried out by applying to $\boldsymbol{\mathcal{Z}}$ a suitable loss function, here denoted with the symbol $\mathcal{L}(\cdot)$. The optimal vector of forecasts is thus the element $\boldsymbol{z}^* \in \boldsymbol{\mathcal{Z}}$ minimizing this function, i.e. $\boldsymbol{z}^* = \min \mathcal{L}(\boldsymbol{\mathcal{Z}})$. This optimality

condition is expressed as

$$\mathbf{z}^* = f(\mu^*, \delta^*), \tag{9}$$

being the arguments of f respectively the "best" forecasting model and forecast combination technique. This last step, by ruling out the less performing combination method(s), has been introduced in order to reduce the overall uncertainty level of the analysis. In fact, suppose that the original set $\mathcal{D}$ reduces to $\mathcal{D}'$—being clearly $|\mathcal{D}'| < |\mathcal{D}|$—the additional amounts of undesired fluctuations and noise—which one can reasonably expect as a consequence of the employment of one (more) under-performing combination method(s)—are avoided. Finally, the model bias β^* is empirically estimated using the in–sample residuals generated by employing the winners techniques μ^* and δ^*, according to an optimal choice made on a set of suitable central tendency functions (fourth optimization).

For the sake of clarity, a more schematic description of the method is given below, in the form of algorithm presented in a step-by-step fashion (Fig. 2).

The proposed method is basically captured by steps 5–7 and 8. For the sake of a more operational comprehension, step 5–7 are now discussed using the matrix notation whereas step 8 will be detailed in Sect. 4.1. Step 5 indicates that once a number M of different forecasts, generated by M models, become available for each level of the hierarchy, they are reconcilied one at a time through $\mathcal{R}$. In practice, the reconciliation function $\mathcal{R}$, applied to the given hierarchical structure, generates a sequence of $H \times 1$ vectors of reconciled predictions $\mathbf{p}_j^H$, as below schematized (Fig. 3).

Each vector $\mathbf{p}_{H,j}$ can be seen as the column of a matrix, say $\mathbf{P}_{H,M}$, containing all the M model-dependent reconcilied forecasts. In the following step 6, these M (column)-vectors of predictions are combined according to a number D of different combination methods $(\delta_1, \ldots, \delta_D)$. In essence, they are sequencially and exhaustively applied to each of the possible combinations of order $\{k = 1, \ldots, \text{ncol}(\boldsymbol{P})\}$ of the column vectors of $\mathbf{P}_{H,M}$. Defining the combination (without repetition) function with C_k and setting, for instance, the combination order $k = k_0 < M$, the submatrix $\mathbf{P}_{H,M}^{k_0} = C_{k_0}(\mathbf{P}_{H,M})$ stores all the $\binom{M}{k_0}$ combinations of the forecasts $\mathbf{p}_j^H;\ j = 1, \ldots, \text{ncol}(\mathbf{P}_{H,M}^{k_0})$, called $\boldsymbol{z}_j^H$, as illustrated in the Fig. 4.

By looping over all of the combination orders $k = 1, \ldots, \text{ncol}(\mathbf{P}_{H,M})$, the matrix $\boldsymbol{Z}$ containing all the possible combination of the M model dependent reconcilied forecasts is obtained. This matrix is called $\boldsymbol{Z}$ and has dimensions

$$H \times D * \text{ncol}(\mathbf{P}_{H,M}). \tag{10}$$

Step 7 translate into simply applying a suitable loss function to $\boldsymbol{Z}$ (column-wise), until the final vector of predictions $\boldsymbol{z}^*$, verifying the minimum condition $\min \mathcal{L}(\mathbf{Z})$, is extracted.

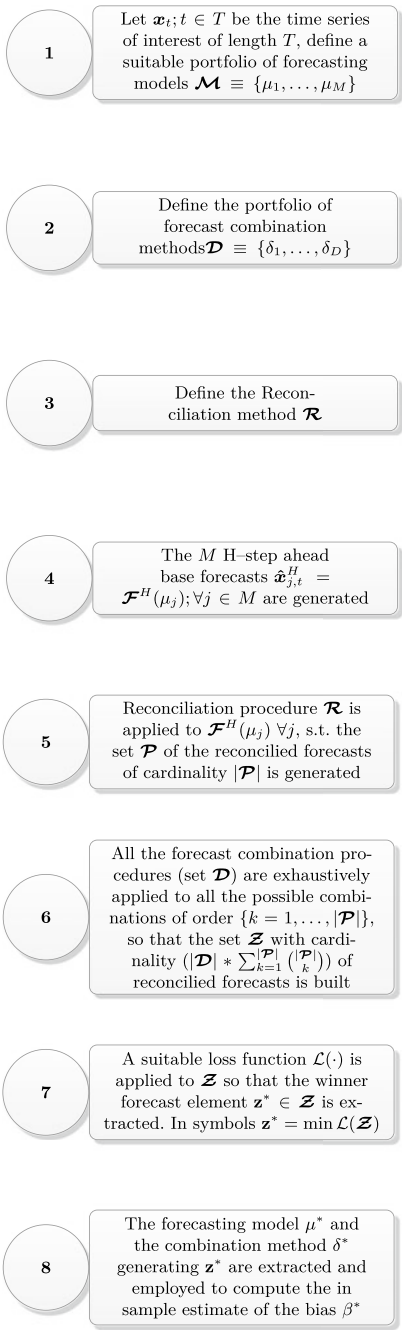


Fig. 2 Algorithm of the proposed method

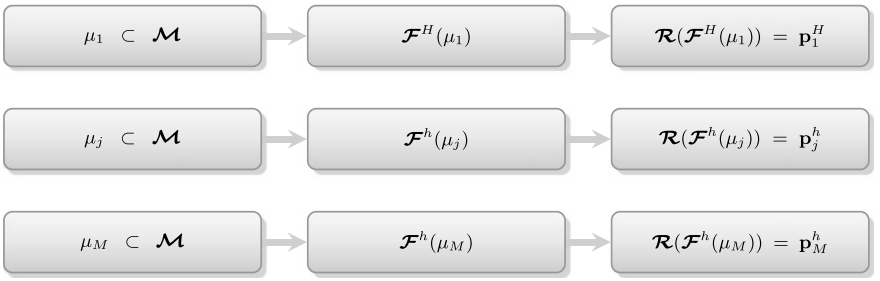


Fig. 3 Illustration of the procedure to obtain the vectors of reconciled forecasts

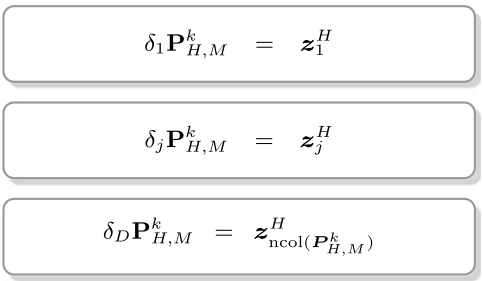


Fig. 4 Illustration of the procedure to obtain the vectors of model dependent reconcilied forecasts

4.1 The Bias Correction Procedure (Step 8)

It is well known that a perfectly unbiased forecast is a condition not frequently met in many real-life applications. Unfortunately, the unbiasedness of the forecasts reconciliation method chosen (Eq. 2), depends on the unbiasedness of each and every base forecasts, as proved by [39]. The proposed method can alleviate this problem as it is designed to generate a "big" competition set ($\mathcal{Z}$), made up with more "balanced" forecasts (thanks to the forecast combinations techniques applied) and thus more likely to perform better than methods generating fewer or just one forecast vector.

The bias correction of the forecast values stored in the vector $\mathbf{z}^*$—obtained in step 7 of Fig. 1—is performed using an improved version of the simple, yet effective, procedure discussed in [34]. In more details, the adopted method translates into a six–step iterative procedure, designed to empirically estimate a set of in-sample tentative biases $\{\boldsymbol{\beta} \equiv \beta_1, \beta_2, \ldots, \beta_B\}$, each of them obtained according to a predefined, arbitrary set of suitable central tendency functions $\{a_1, a_2, \ldots, a_A \subset \boldsymbol{\mathcal{A}}\}$, being $|\boldsymbol{\beta}| = |\boldsymbol{\mathcal{A}}|$ (or, equivalently, $B = A$).

Recalling that $\boldsymbol{x}_t$ is the observed time series, let us denote with $\hat{\boldsymbol{x}}_t^* = f(\mu^*, \delta^*)$ its one-step-ahead predictions—obtained according to the optimal forecasting model and prediction combination method (see Eq. 9)—and with $\boldsymbol{\varepsilon_t}|\beta_j$ the vector of residuals between these two series conditional to a given central tendency function, i.e. $\boldsymbol{\varepsilon_t}|\beta_j = \boldsymbol{x}_t - \hat{\boldsymbol{x}}_t^*|\beta_j(\alpha_j)$. In what follows, the term α_j is omitted as it is under-

stood the dependency relationship between bias and central tendency function, i.e. bias $= \beta_j(\alpha_j)$. The set $\boldsymbol{\beta}$ is thus generated by iteratively and exhaustively applying each function in $\boldsymbol{\mathcal{A}}$ to the bias adjusting equations which, once expressed in terms of residuals, read as follows:

$$^{\beta_j}\boldsymbol{\varepsilon_t} = \boldsymbol{x}_t - \hat{\boldsymbol{x}}_t^*|\beta_j; \quad j = 1, 2, \ldots, B, \tag{11}$$

$$^{\beta_j}\boldsymbol{\eta_t} = \frac{\boldsymbol{x}_t}{\hat{\boldsymbol{x}}_t^*|\beta_j}; \quad j = 1, 2, \ldots, B. \tag{12}$$

Equations 11–12 differ only for that the one-step ahead predictions are respectively adjusted additively and multiplicatively. Finally, by applying a suitable loss function, $\mathcal{E}(\cdot)$ to each of the vectors $^{\beta_j}\boldsymbol{\varepsilon_t}$ and $^{\beta_j}\boldsymbol{\eta_t}$, the optimal bias estimation is its minimzer, i.e.

$$\beta^* = \min_{\mathcal{E}}(^{\beta_j}\boldsymbol{\varepsilon_t}; ^{\beta_j}\boldsymbol{\eta}); \quad j = 1, 2, \ldots, B. \tag{13}$$

Once the final bias is computed, it can be readily applied in a forward looking fashion, i.e.

$$^{a}\boldsymbol{y_{t,H}} = \mathbf{z}_{,H}^* + \beta^*(\boldsymbol{x}_t - \mathbf{z}_t^*) \tag{14}$$

or

$$^{m}\boldsymbol{y_{t,H}} = \mathbf{z}_{,H}^* * \beta^* \frac{\boldsymbol{x}_t}{\mathbf{z}_t^*}, \tag{15}$$

according to whether the winner central tendency function is applied in an additive (Eq. 14) or multiplicative fashion (Eq. 15). The generalized notations for the first term in equations is

$$^{u}\boldsymbol{y_{t,H}}, \tag{16}$$

which represent the final predictor. In such an approach, the future and the past are assumed to be affected by the same amount of bias. Such an assumption, under stationaity of the observed time series and a "sufficient" sample size, might not be considered unreasonable.

In the case of $\{\boldsymbol{\beta} \equiv \beta_1\}$ with β_1 the mean function, Eqs. 14–15 respectively are as follows:

$$^{a}\boldsymbol{y_{t,H}} = \mathbf{z}_{t,H}^* + \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}_t - \mathbf{z}_t^*\right)$$

or

$$^{m}\boldsymbol{y_{t,H}} = \mathbf{z}_{t,H}^* * \frac{1}{T}\sum_{t=1}^{T}\frac{\boldsymbol{x}_t}{\mathbf{z}_t^*},$$

and thus equivalent (saved for the notation) to the procedure discussed in [34] (Eqs. 3–4), page 21.

In the empirical application of Sect. 6, the central tendency functions applied are: mean, median and root mean square, respectively denoted by the lowercase Greek letters π, τ and ξ. Their mathematical representations are as follows:

$$\pi(\cdot) = \frac{1}{T}\sum_{t=1}^{T} x_t, \qquad \tau(\cdot) = \frac{1}{2}x_{(\lfloor T+1)/2\rfloor} + x_{\lceil (T+1)/2\rceil}, \qquad \xi(\cdot) = \frac{1}{T}\sqrt{\sum_{t=1}^{T} x_t^2}. \tag{17}$$

In the case of the median ($\tau(\cdot)$) x is an ordered list of T values, and the symbols $\lfloor\cdot\rfloor$ $\lceil\cdot\rceil$ denote the floor and ceiling functions, respectively.

5 Justification of the Method

The effectiveness of the proposed method is, in general, conditioned to the choices of the statistical prediction models and the forecast combination techniques included in the sets $\mathcal{M}$ and $\mathcal{D}$, as well as to the selection of the most suitable central tendency functions (the set $\mathcal{A}$). A careful building of those sets (our multidimensional search space), is a prerequisite for the proposed method to properly perform. Its final dimensions are as in Expression (Eq. 10) plus twice the number of central density functions considered, used in both additive and multiplicative fashion, i.e.

$$H \times \left[2 * |\mathcal{A}| + D * \text{ncol}(\mathbf{P}_{H,M})\right].$$

The point of strength of the method is thus related to the availability of a potentially large number of multiple choices, all of them derived in a multiple combination set-up (in terms of statistical prediction model, forecast combination and bias correction), so that the selected forecast vector are the minimizer of a bi-dimensional loss function $(\mathcal{L}, \mathcal{E})$. In addition, the method is very flexible, as it can work with all the methods deemed suitable for the problem at hand, being the only limit the computational time. This is certainly an issue, which, however, can be easily circumvented thanks to the structure of the method itself, which is naturally prone to be parallelized.

According to the bias-variance decomposition approach, the mean square error (MSE) can be decomposed into a bias β and a variance (V) terms, i.e.

$$MSE = \beta^2 + V \tag{18}$$

In what follows, the advantages related to the proposed method will be illustrated in terms of Eq. 18. Firstly, it is noted how, in general, simpler models tend to produce large biases and small variances whereas complex models behave in the opposite way. The proposed method is designed to overcome such an issue not only because it can employ a combination of several models with different levels of complexity

but also because it selects the "best" combination technique (stored in the set $\mathcal{D}$) according to the data set under investigation. This last feature is clearly a plus, since there is not such combination techniques able to perform optimally in any circumstances. For example, [31] found that there are cases where a simple average combination may be more robust than weighted average combinations. Therefore, by iteratively testing many different techniques, one is more likely to find the most suitable (if not the "optimal") one.

Bias-wise, the advantages of this method are related to its self-balancing and self-adjusting features, the former being induced by the bias compensation phenomenon, more likely to occur in a multi-model set up, whereas the latter relies on the bias correction procedure, given in Sect. 4.1. In particular, its effectiveness in bias reduction is related to the fact that the self-adjustment procedure uses a vector of forecast which, by design, has already been controlled for bias. To see this, let us express the generic forecast combination $\tilde{\boldsymbol{x}}_t^H$ as

$$\tilde{\boldsymbol{x}}_t^H = \sum_{i=1}^{\boldsymbol{M}} w_i \boldsymbol{\mathcal{F}}^H(\mu_i),$$

with $w_i = f(\tilde{\delta})$ the combination weights generated by the combination method $\tilde{\delta}$ and $\boldsymbol{\mathcal{F}}^H(\mu_j)$; $j = 1, \ldots, M$ are base forecasts generated by the M statistical models entertained (see Fig. 1 step 1 and 4). Assuming $0 \leq w_j \leq 1$, $\sum_{i=1}^{M} w_i = 1$ and the vector of "future" observations of length H to be known, the total amount of bias of the combined forecast is given by

$$\begin{aligned} \beta &= \mathbf{E}(\tilde{\boldsymbol{x}}_t^H - \boldsymbol{x}_t^H) \\ &= \sum_{i=1}^{M} w_i [\mathbf{E}\tilde{\boldsymbol{x}}_{t,i}^H - \mathbf{E}\boldsymbol{x}_t^H] \\ &= \sum_{i=1}^{M} w_i \beta_i, \end{aligned} \tag{19}$$

where the subscript i is used to refer to a specific model, in terms of generated bias (β_i) and forecast ($\tilde{\boldsymbol{x}}_{t,i}^H$). The right term of Eq. 19 shows that the bias of the forecast combination is the weighted average of the biases of the base forecasts and thus, provided that their magnitude is comparable, one can reasonably expect an overall bias reduction due to cancellation effects. Such a phenomenon is not rare, since—in general—it is not common for all the biases to show the same sign. Since the bias-correcting method—discussed in Sect. 4.1—is applied on an already optimally combined vector of forecasts $\mathbf{z}_t^* = \boldsymbol{x}_t(\delta^*, \mu^*)$, a less pronounced bias can be expected in the final predictions, given by

$${}^{u}\boldsymbol{y}_{t,\boldsymbol{H}} = \boldsymbol{z}_{t,H}^{*} + \boldsymbol{\beta}^{*}(\boldsymbol{x}_t - \boldsymbol{z}_t^{*}), \tag{20}$$

where β^* is as in Eq. 13.

The proposed method can also help keep low the variance of the forecasts in an amount inversely proportional to the correlation coefficients computed between the competing forecasts and proportional to the reduction in the standard errors induced by the reconciliation procedure adopted $\mathcal{R}$. To see this, denoting with σ_i^2 the variance of the individual forecast i and with $\gamma_{i,j}$ the correlation coefficient computed on a generic pair (i, j) of forecasts, we use the following inequality (derived in [3]), i.e.

$$V \leq \sum_{i=1}^{N} w_i \sigma_i^2 - 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j (1 - \gamma_{i,j}) \sigma_i \sigma_j, \tag{21}$$

where V is the variance of the combination of forecasts. Inequality (Eq. 21) states that V tends to be considerably less than the average of the individual forecasts in an amount depending on $\frac{1}{\gamma_{i,j}}$, meaning that less correlated forecasts are beneficial in terms of variance reduction. Such a situation is more likely to occur in procedures which, as the one proposed, grant a "sufficient" number of predictions. However, there is another point in favor of the proposed method on this matter. In fact, recalling that the different combinations of the forecasts are performed on already reconciled vectors of predictions $\mathbf{z}_t^H$—according to the adopted procedure $\mathcal{R}$—by virtue of Eq. 4 their variances obey to the following inequality:

$$Var(\mathbf{z}_t^H) \; < Var(\boldsymbol{x}_t^H); \forall \mathbf{z} \in \boldsymbol{\mathcal{Z}}. \tag{22}$$

Therefore, the overall level of variance in Eq. 21 decreases of an amount inversely proportional to $Var(\mathbf{z}_t^H)$.

6 Empirical Study

In this section the goodness of the proposed method will be evaluated using the official time series related to the number of people tested positive to the SARS-CoV-2 in each of the Italian regions, between February 24th 2020 and October 7th 2020. The whole data set—issued by the Italian National Institute of Health—are publicly and freely available at the web address https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni. The data, sampled at a daily frequency, are stored in a matrix called $\boldsymbol{\mathcal{O}}$ (see Table 1) of dimension 227×21, where 21 are the Italian regions. From a strictly administrative point of view, the number of the Italian regions amounts to 20, however, for one of them, called Trentino Alto Adige, the data are split according to its two main provinces: Trento and Bolzano. As reported in the same Table 1, the proposed procedure is trained on a portion of the data matrix, called $\boldsymbol{\mathcal{O}}^{train}$, of dimensions 197×21 and time span February 24th—September 7th, whereas the test part is carried out on a set, called $\boldsymbol{\mathcal{O}}^{test}$, whose dimensions are $H = 30 \times 21$ (the time span is from September 8th to October 7th). Finally, 30 days ahead "real-life" estimates—in the sense that they are related to future values which are unknown at

Table 1 The employed data set and its portions defined according to the different purposes served

Symbol	Start date	End date	Sample size
$\mathcal{O}$	February 24th	October 7th	227
$\mathcal{O}^{train}$	February 24th	September 7th	197
$\mathcal{O}^{test}$	September 8th	October 7th	30
$\mathcal{O}^{fore}$	October 8th	November 7th	30

Table 2 Symbols employed to identify the statistical models

Model (μ)	Symbol
ARFIMA	A
ETS	E
THETA	T
ARIMA	B

Table 3 Symbols employed to identify the central tendency functions

Method (δ)	Symbol
Ordinary least square	OLS
Least absolute deviation	LAD
Newbold and granger	NG
Simple average	SA

the time of their computations—for the time window October 8th—November 7th will be stored in the matrix $\mathcal{O}^{fore}$. Since the proposed procedure combines a number of models (μ's), combination methods (δ's) and central tendency functions ($\mathcal{E}'s$), for each of those, conventional symbols are respectively given in Tables 2, 3 and 4. Consistently with the convention introduced in Eq. 16, in Table 4 the superscript u is used to indicate the type of bias considered, i.e. additive ($u = a$) or multiplicative ($u = m$). Finally, to efficiently keep track of the outcomes of the method, in Table 5 the whole set of model combinations employed—respectively of class $k = 4, 3, 2$—are provided. Since we have four different combination methods, each of the 11 model combination (reported in Table 5) are performed four times, which yields a total of 44 method-dependent combinations. One of them, for example, is the forecasts combination generated by combining an ARFIMA and an ETS models using the method Ordinary Least Squares. This information is conveniently conveyed by the symbol *OLS–AE*.

Recalling that with ${}^{u}\boldsymbol{y}_{t,H}$ the final predictions yielded by the proposed method are denoted (see Eq. 16), the loss function employed ($\mathcal{L}$) is the Root Mean Square Error (RMSE), given by $\sqrt{\frac{1}{T}\sum_{h=1}^{30}(\boldsymbol{x}_{t,H} - {}^{u}\boldsymbol{y}_{t,H})^2}$. The same function is adopted

Table 4 Symbols employed to identify the statistical models

Central tendency function	Symbol
Mean	${}^{u}\pi$
Median	${}^{u}\tau$
RMS	${}^{u}\xi$

Table 5 Combinations of models of class $K = 4, 3, 2$ attempted for each of the four model-dependent reconcilied forecasts

Number	K	Model combination
1	4	A–E–T–B
2	3	A–B–E
3		A–B–T
4		A–E–T
5		B–E–T
6	2	A–B
7		A–E
8		A–T
9		B–E
10		B–T
11		E–T

in–sample to select the best 3–tuple $(\mu^*, \delta^*, \mathcal{E}^*)$, i.e. $\sqrt{\frac{1}{T}\sum_{t=1}^{227}(\boldsymbol{x_t} - {}^{u}\boldsymbol{y_t})^2}$ and to evaluate the method's performance in the test set $\mathcal{O}^{test}$. Finally, the out of sample estimate of the bias, in the sequel denoted by the symbol β^{out}, has been computed on the set $\mathcal{O}^{test}$, using the formula $\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}_t - {}^{u}\boldsymbol{y}_{\boldsymbol{t},\boldsymbol{H}}\right)$.

6.1 Performances of the Method

The performances of the method are summarized in Appendix A and in Table 6. In particular, in Appendix A the observations belonging to the test set $\mathcal{O}^{test}$ (black line) and the related estimates ${}^{u}\boldsymbol{y}_{\boldsymbol{t},\boldsymbol{H}}$ (red line) are depicted for some Italian regions. The first three columns of Table 6 respectively indicate the name of the Italian regions, the winner combination (μ^*, δ^*) and the related RMSE values $(\mathcal{L}^*)$. The best central tendency function $(\mathcal{E}^*)$, the selected bias corrector (β^*) and the estimated out-of-sample bias (β^{out}) are given in columns four, five, six. In the last two columns the $RMSE$ values $(\mathcal{L})$ relative to each of the forecasting models (reported in Table 2) taken separately, are recorded. The accuracy of the proposed method seems to be very good as, in almost all the cases, the winner combinations deliver better predictions than all the statistical models singularly considered and, in many cases, outperform

them. This is the case, for example, of the Campania and Calabria regions. Here (see Table 6), the recorded $RMSE$ is respectively equal to 34.7 and 63.5, far below the values obtained using the best statistical models, i.e. $Theta$ and ETS, which respectively scored $\mathcal{L} = 143.6$ and 265.1. In addition, the proposed procedure shows always from very low to negligible in-sample amounts of bias. The most selected central tendency functions is ${}^{a}\tau$ and ${}^{a}\pi$, whereas it is noted that the RMS function (${}^{u}\xi$) has never been chosen. Regarding the out-of-sample bias, if on one hand, as expected, it is always $\beta^{out} > \beta^{*}$, on the other hand it can be said that the magnitude of the values assumed by β^{out} can be deemed acceptable. This is especially true if one considers the length of the prediction window ($H = 30$ days) compared with the available data set. In particular, six regions (Valle d'Aosta, Trento, Molise, Lazio, Abruzzo, Calabria) show interesting values for the out–of–sample bias, being $\beta^{out} < 10$. The worse performances of the method refers to the region of Sardegna, where the consistent underestimations of the true values lead to a recorded bias of around 504. This fact can be explained by looking at the irregular, bumpy shape of this time series, reported in Appendix B (see the plot related to Sardegna), which might have introduced distortions in the model estimators. In Appendix B, the graphical results of a pure out-of-sample application of each of the winner combinations are reported. In more details, the region-specific winner combinations (μ^{*}, δ^{*}) are applied to the whole set $\mathcal{O}$, so that the $H = 30$ days–ahead resulting forecasts—stored in the set $\mathcal{O}^{fore}$—are the pure forecasts for the period October 8th—November 7th. In this Appendix, the regional time series in $\mathcal{O}$ (true observations) are plotted in black whereas the predictions are in red color. The analysis of these Figures suggest a slower acceleration in the growth of positive cases in some of the north regions (e.g. Lombardia, Trento, Liguria) whereas the center and south regions might face a strong increase of positives. This seems likely to happen in the Campania, Basilicata and Molise regions. The number of positive for Italy, predicted for the end of the pure forecast period (November 7th), is about of 140,000.

7 Conclusions and the Future Directions

The present paper provides sufficient evidences that reconciliation serves the double purpose of generating coherent forecast with improved accuracy, under a multidimensional optimization constraint. The proposed method is designed to handle the increased amount of uncertainty surrounding the forecasting, as one carries out the prediction exercise at a progressively more disaggregated levels. The novelty of this procedure is the application of the forecast combination techniques to already optimally combined forecasts, generated by different prediction models. In this regard, the proposed procedure can be improved by increasing the portfolio of the statistical models entertained under a suitable program architecture. The second line of future research refers to the bias estimation, whose uncertainty might estimated using a suitable resampling scheme.

Table 6 Performances of the method for each of the Italian regions. Outcomes of the winner models and of the single statistical models. See text for details

Region	Winner combination	$\mathcal{L}^*$	$\mathcal{E}^*$	β^*	β^{out}	Single models	$\mathcal{L}$
Piemonte	$LAD-BET$	135.6	${}^{a}\tau$	≈ 0	54.09	$ARFIMA$	1050.8
						ETS	1065.8
						θ	1271.8
						$ARIMA$	683.6
Val d'Aosta	$SA-BT$	8.6	${}^{a}\tau$	≈ 0	-2.25	$ARFIMA$	60.7
						ETS	177.5
						θ	30.5
						$ARIMA$	29.5
Lombardia	$SA-ET$	172.6	${}^{m}\tau$	1.0	70.83	$ARFIMA$	3184.0
						ETS	911.7
						θ	966.1
						$ARIMA$	1713.8
Bolzano	$OLS-ET$	223.2	${}^{a}\pi$	≈ 0	190.9	$ARFIMA$	468.9
						ETS	227.6
						θ	271.2
						$ARIMA$	248.8
Trento	$NG-BE$	29.4	${}^{a}\tau$	-0.13	8.95	$ARFIMA$	113.7
						ETS	30.0
						θ	238.2
						$ARIMA$	211.5
Veneto	$NG-BE$	58.5	${}^{a}\pi$	0.55	49.01	$ARFIMA$	620.1
						ETS	95.5
						θ	259.9
						$ARIMA$	49.6
Friuli Venezia Giulia	$SA-BE$	65.9	${}^{a}\pi$	0.70	18.08	$ARFIMA$	475.4
						ETS	140.6
						θ	763.5
						$ARIMA$	155.6
Liguria	$LAD-ET$	124.2	${}^{a}\tau$	≈ 0	-28.13	$ARFIMA$	123.8
						ETS	126.8
						θ	1057.7
						$ARIMA$	241.9
Emilia Romagna	$LAD-BT$	296.7	${}^{a}\tau$	≈ 0	-160.01	$ARFIMA$	2203.4
						ETS	343.5
						θ	724.5
						$ARIMA$	274.9

Table 6 (continued)

Region	Model	$\mathcal{L}^*$	$\mathcal{E}^*$	β^*	α^*	Single models	$\mathcal{L}$
Toscana	$LAD-AB$	259.6	${}^{a}\tau$	≈ 0	149.8	$ARFIMA$	3417.7
						ETS	1202.4
						θ	2485.9
						$ARIMA$	310.0
Umbria	$OLS-AB$	133.3	${}^{a}\pi$	≈ 0	104.54	$ARFIMA$	701.9
						ETS	254.5
						θ	344.4
						$ARIMA$	168.1
Marche	$SA-BET$	211.0	${}^{m}\tau$	1.01	136.65	$ARFIMA$	2191.6
						ETS	320
						θ	1185.7
						$ARIMA$	355.5
Lazio	$SA-ET$	45.9	${}^{a}\tau$	0.13	7.43	$ARFIMA$	522.2
						ETS	153.5
						θ	180.7
						$ARIMA$	64.7
Abruzzo	$LAD-BET$	40.0	${}^{a}\tau$	≈ 0	6.97	$ARFIMA$	740.2
						ETS	70.3
						θ	294.6
						$ARIMA$	40.7
Molise	$SA-AB$	39.2	${}^{a}\pi$	-0.05	-5.8	$ARFIMA$	250.9
						ETS	1229.0
						θ	147.0
						$ARIMA$	265.1
Campania	$SA-BT$	34.7	${}^{a}\tau$	0.05	-30.52	$ARFIMA$	480.7
						ETS	359.8
						θ	143.6
						$ARIMA$	201.3
Puglia	$LAD-BET$	419.4	${}^{a}\tau$	≈ 0	-249.58	$ARFIMA$	1991.5
						ETS	680.6
						θ	2507.5
						$ARIMA$	674.1
Basilicata	$LAD-AB$	75.2	${}^{a}\tau$	≈ 0	-54.68	$ARFIMA$	117.5
						ETS	1453.5
						θ	37.0
						$ARIMA$	417.7

Table 6 (continued)

Region	Model	$\mathcal{L}^*$	$\mathcal{E}^*$	β^*	α^*	Single models	$\mathcal{L}$
Calabria	$OLS - AETB$	63.5	${}^{a}\pi$	≈ 0	4.13	$ARFIMA$	2124.9
						ETS	265.1
						θ	1109.9
						$ARIMA$	302.1
Sicilia	$SA - BET$	70.7	${}^{m}\pi$	0.39	57.68	$ARFIMA$	1436.8
						ETS	333.0
						θ	715.7
						$ARIMA$	195.2
Sardegna	$NG - AB$	581.9	${}^{m}\pi$	-0.05	504.75	$ARFIMA$	983.7
						ETS	998.9
						θ	1204.6
						$ARIMA$	605.4

8 Data Availability

The data that support the findings of this study are openly available in the section "COVID-19/dati-regioni/" at https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni.

9 Disclaimer

The views and opinions expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics.

Appendix A

In this appendix, both the 30 days ahead out–of–sample predictions and the actual data (i.e. the set $\mathcal{O}^{test}$) are plotted for some Italian regions (the remaining are available from the author upon request). The visual inspections of such plots can provide useful information related to the speed at which the virus is spreading. In more details, a good fitting (see, for example, the case of Sicilia) is indicative of more stable dynamics driving the spread of the infection. On the contrary, a poor fitting (see, for example, the case of Sardegna) might denote a rapidly changing situation. Such information can be used to better understand the out–of–sample results provided by the proposed method, as it will be illustrated in Appendix B (Fig. 5).

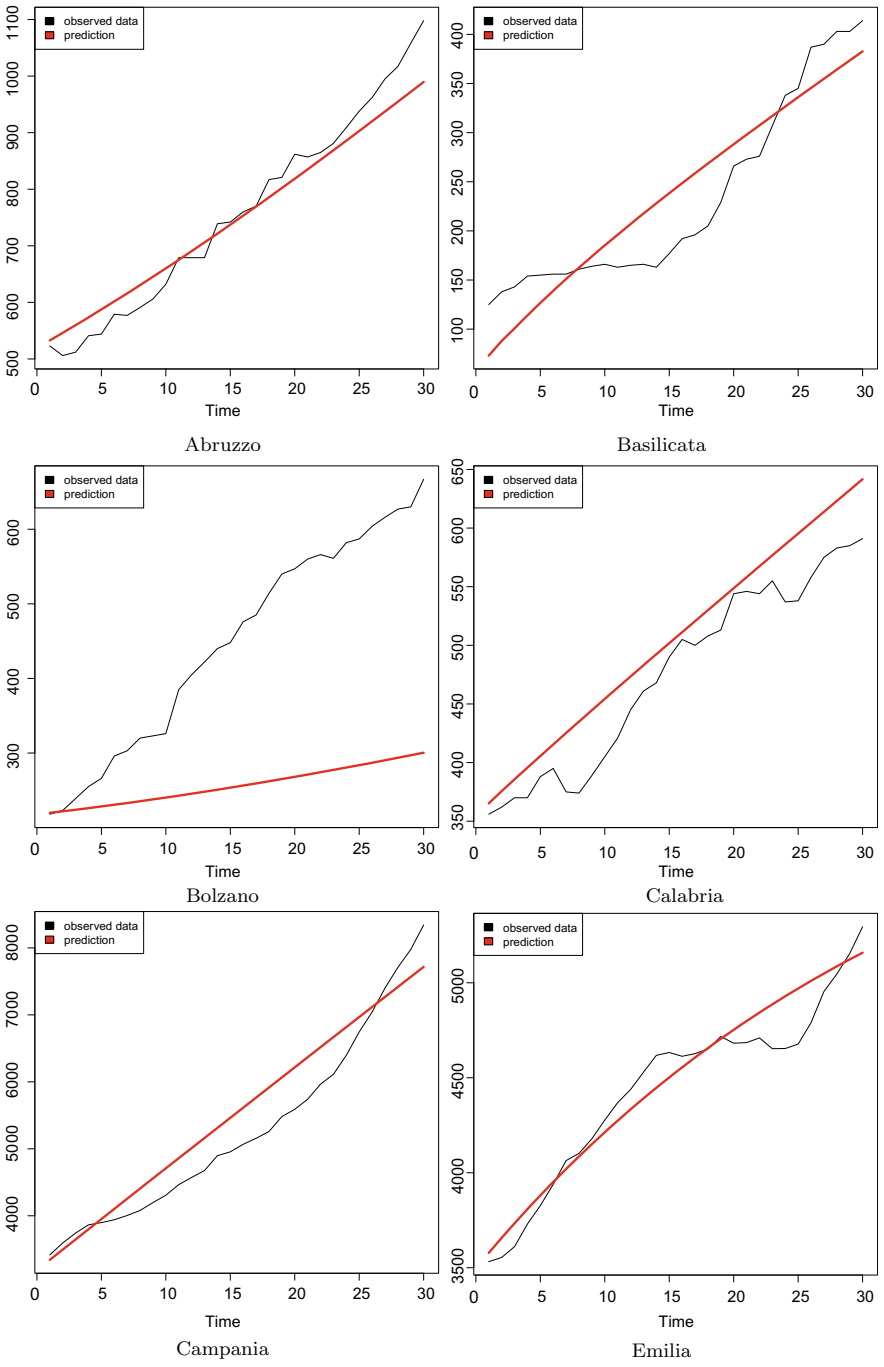


Fig. 5 Out–of–sample predictions (red lines) and actual data (black lines, representing the set $\mathcal{O}^{test}$) related to the Italian regions. Forecast window $H = 30$ days

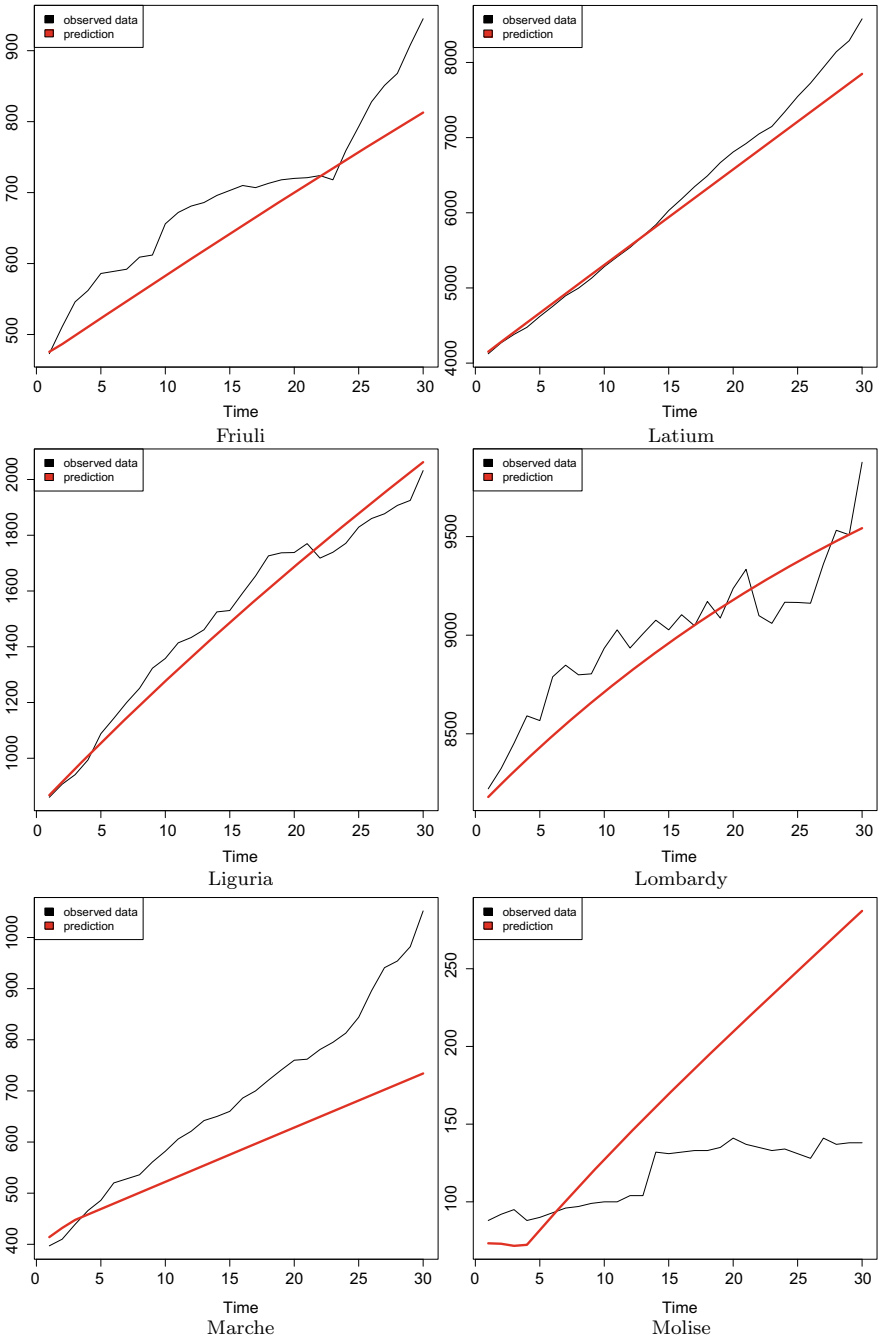


Fig. 5 (continued)

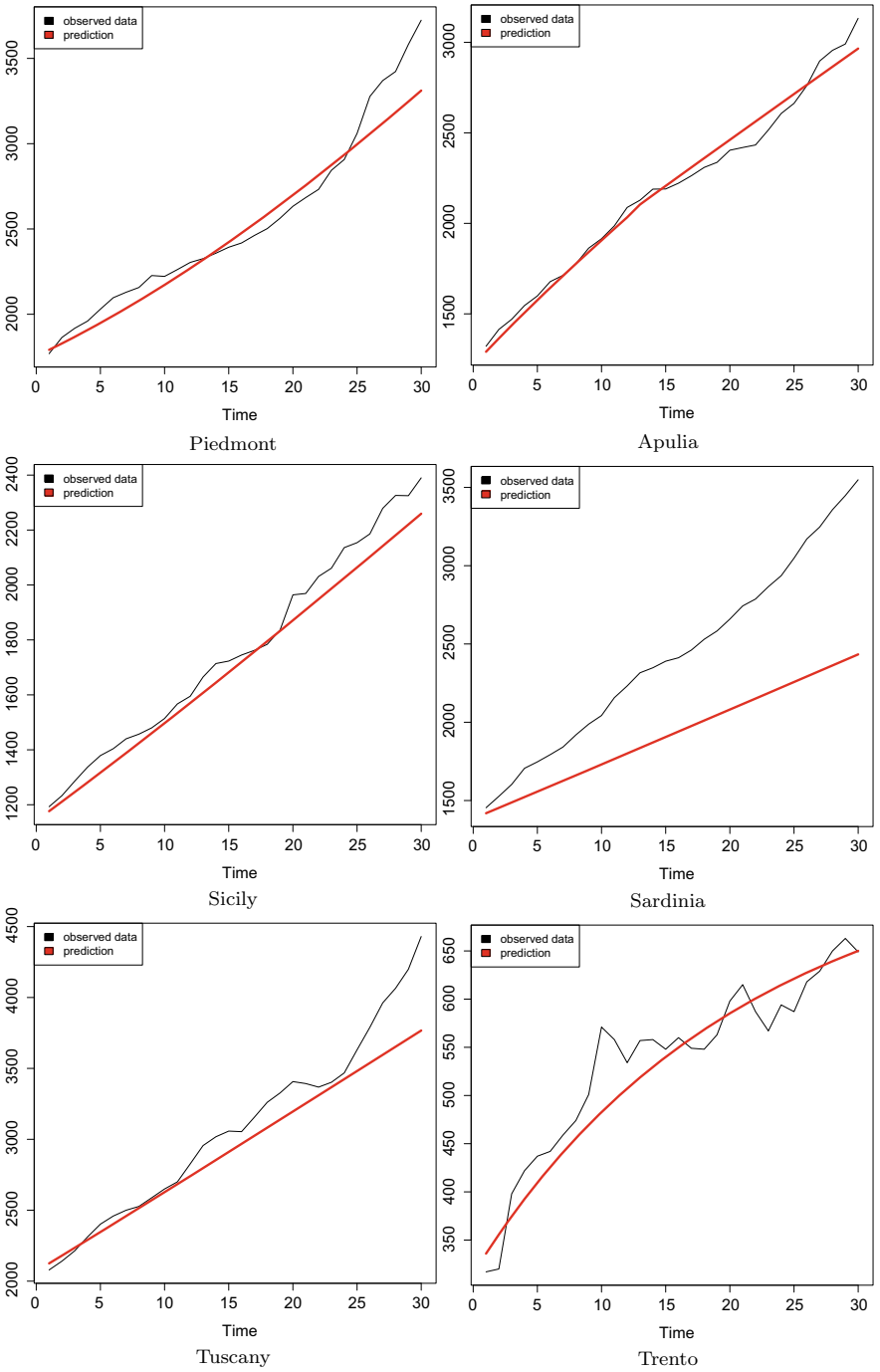


Fig. 5 (continued)

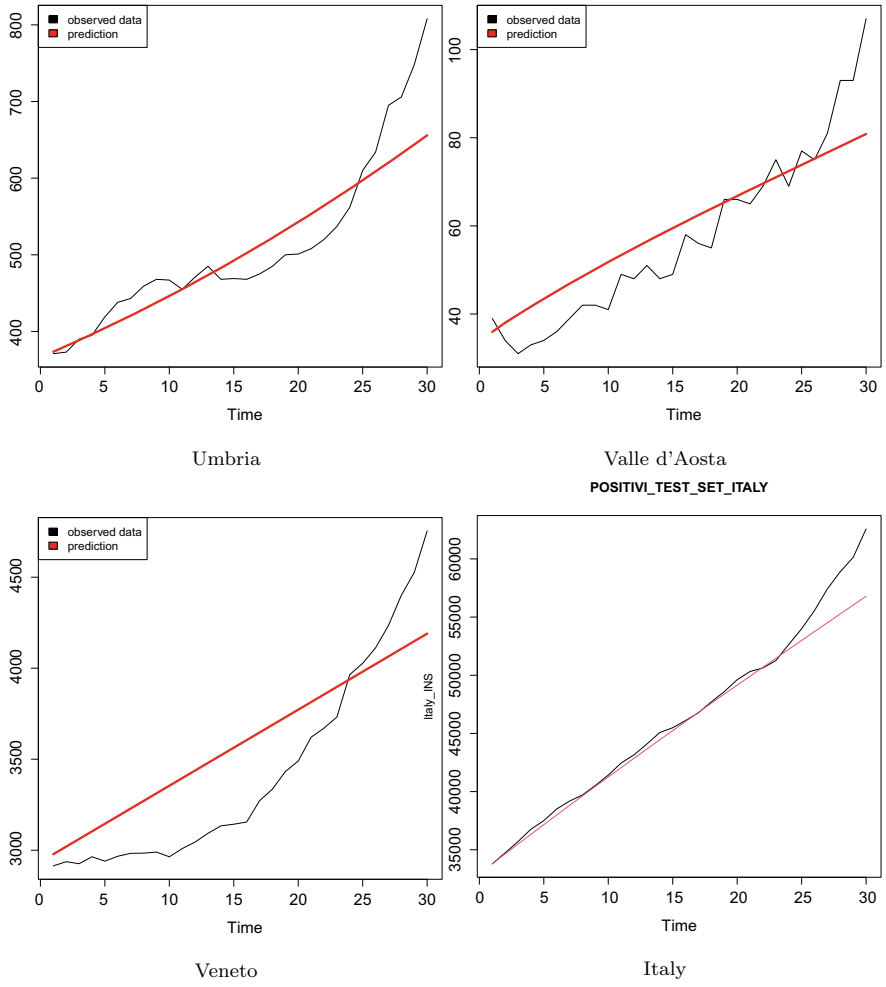


Fig. 5 (continued)

Appendix B

In this appendix, the 30 days ahead out–of–sample pure predictions (i.e. the set $\mathcal{O}^{fore}$, in red) along with the actual data (black lines) are reported for all the Italian regions and Italy (Fig. 6).

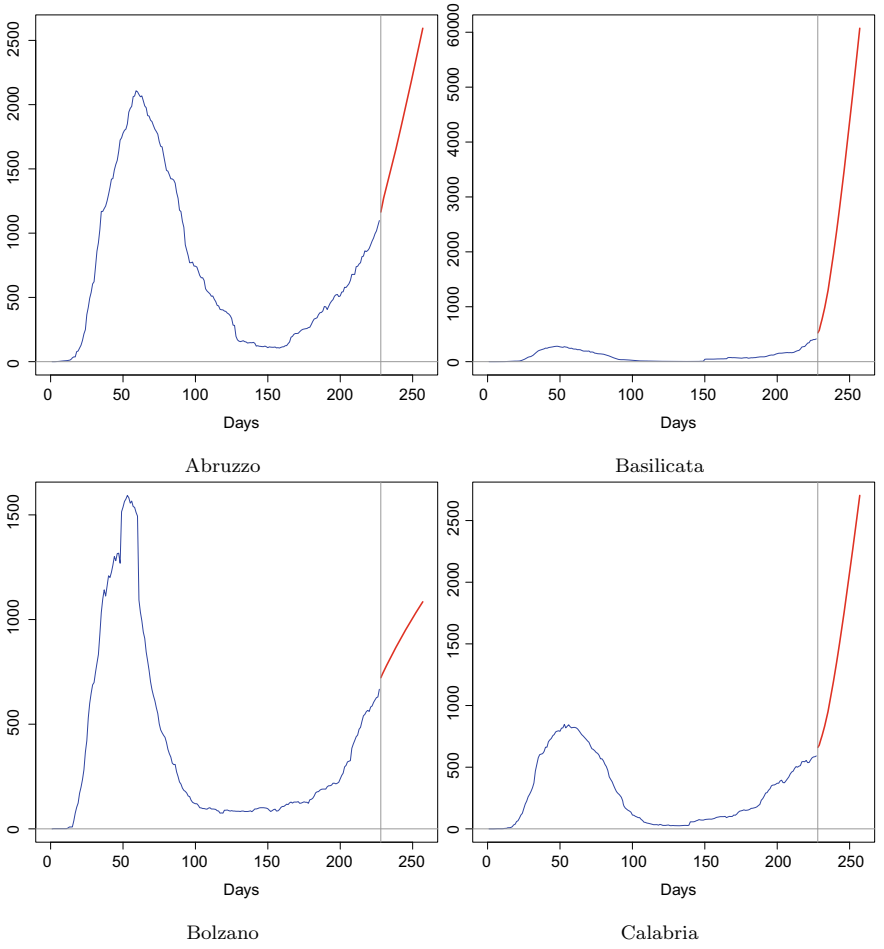


Fig. 6 Out–of–sample pure forecasts. The Black lines refer to the whole set of actual data (the set $\mathcal{O}$) where the predictions for the horizon $H = 3O$ days (the set $\mathcal{O}^{fore}$) are reported in red

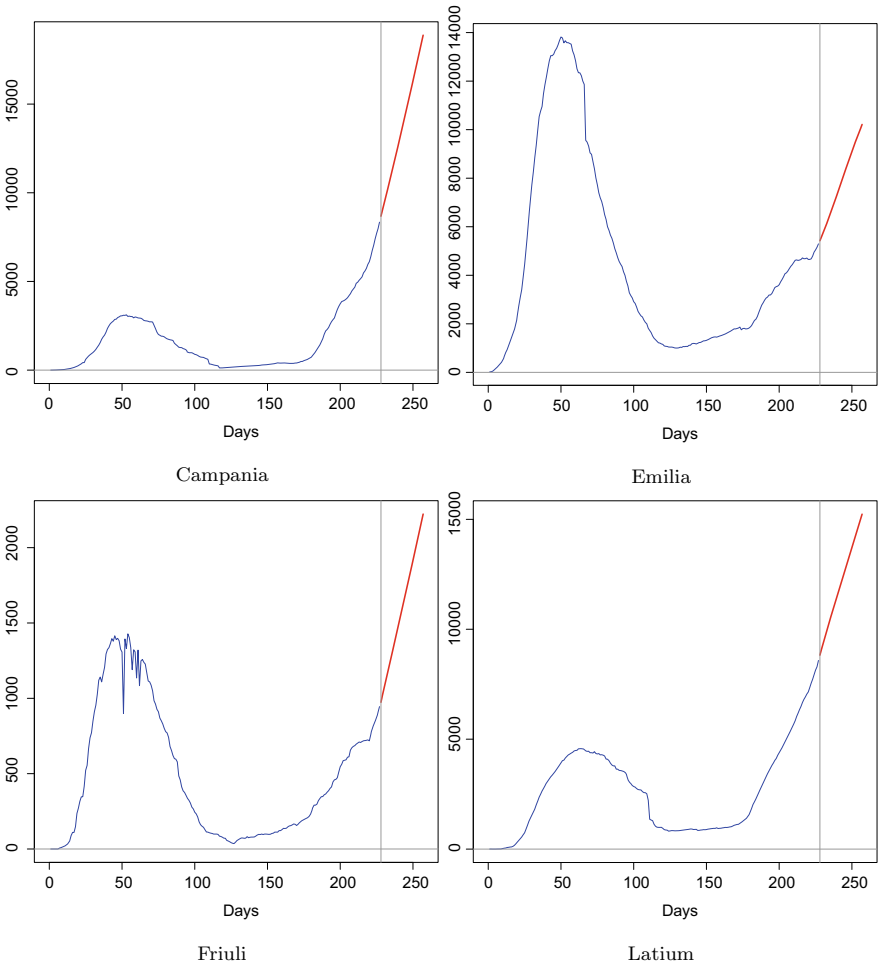


Fig. 6 (continued)

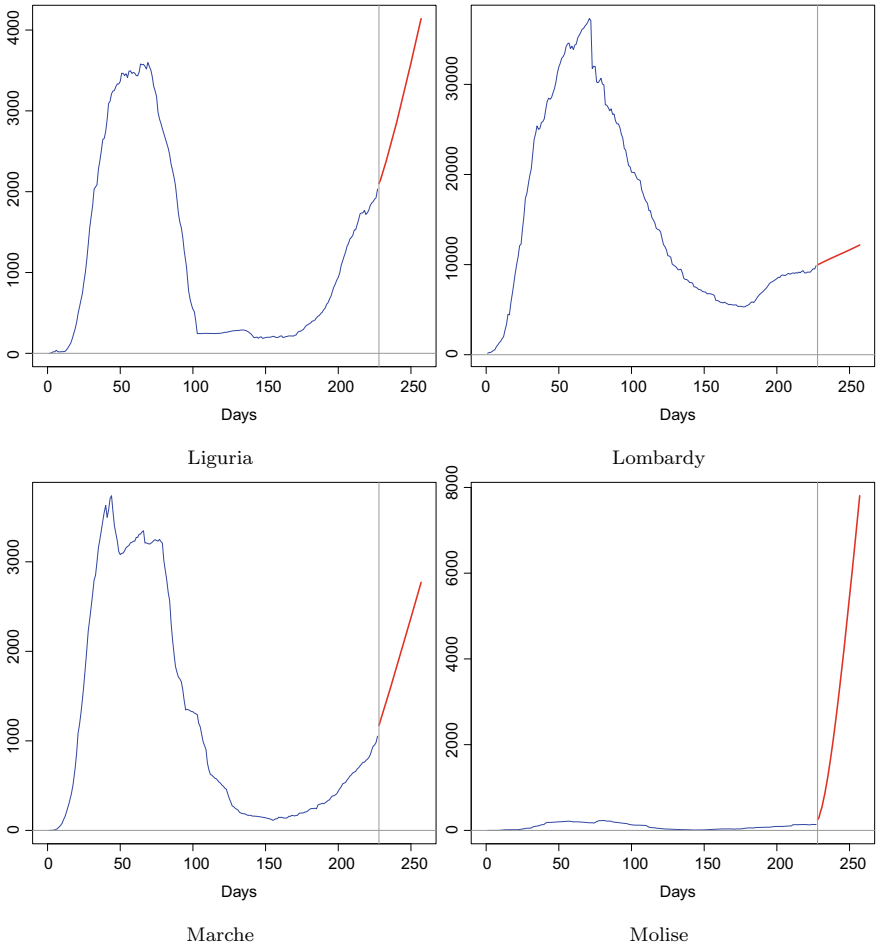


Fig. 6 (continued)

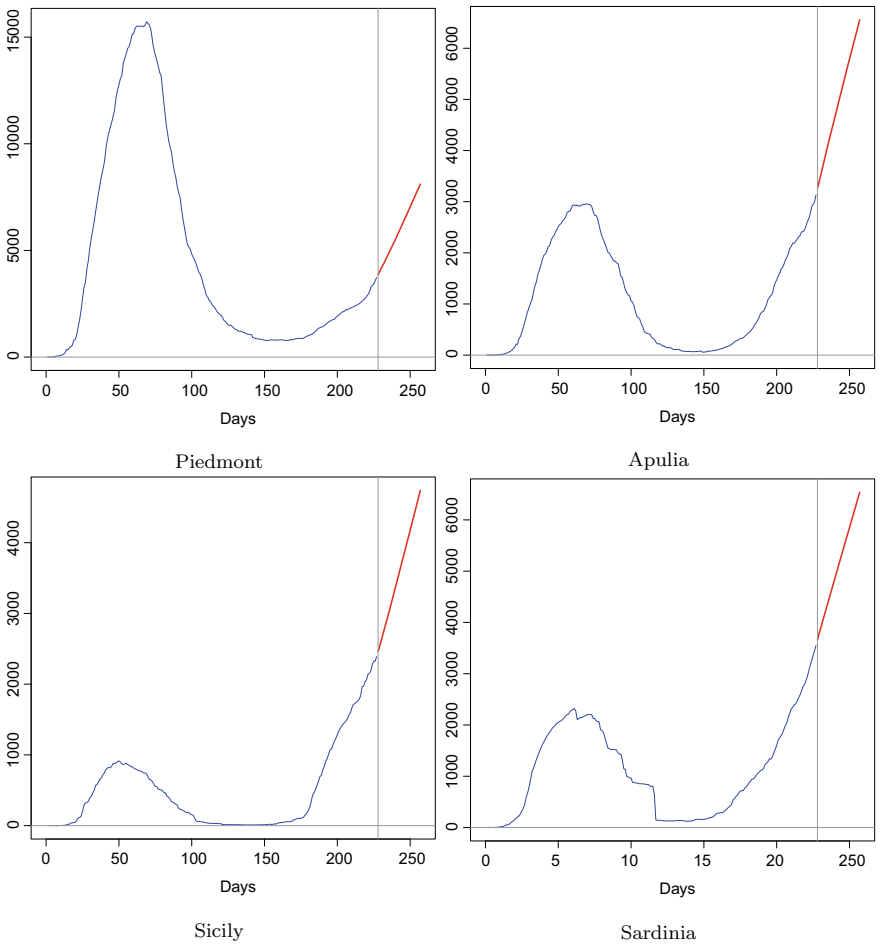


Fig. 6 (continued)

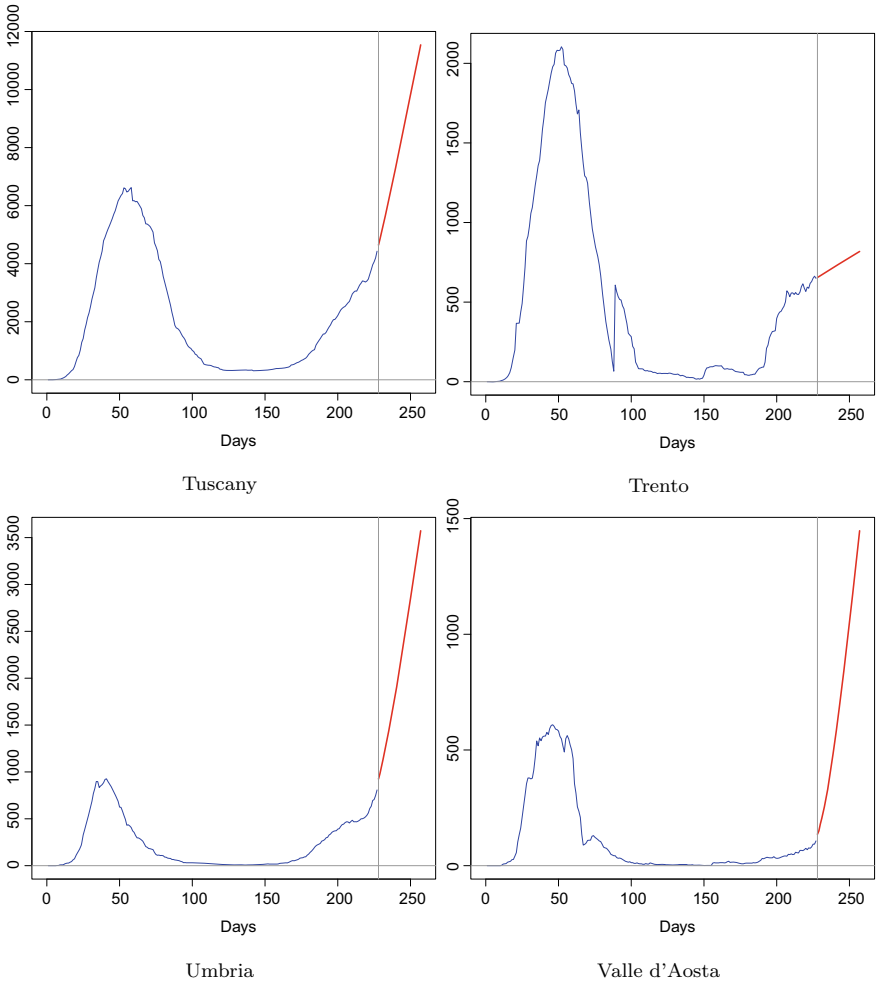


Fig. 6 (continued)

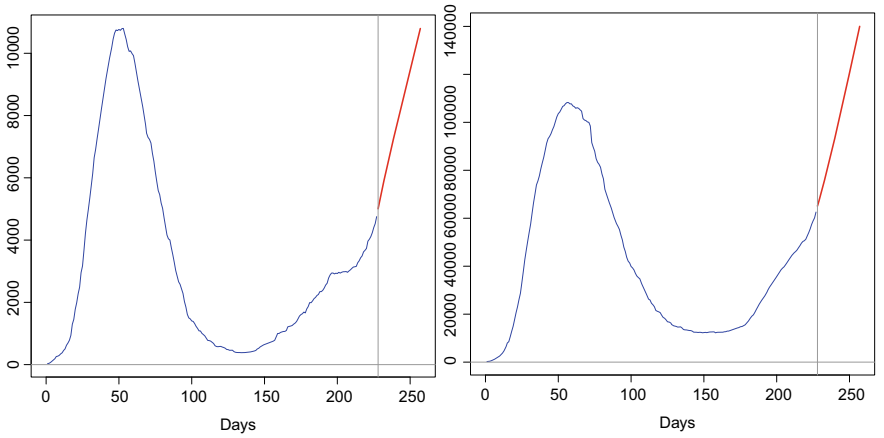


Fig. 6 (continued)

References

1. Assimakopoulos, V., Nikolopoulos, K.: The theta model: a decomposition approach to forecasting. Int. J. Forecast. Int. J. Forecast. **16**(4), 521–530 (2000)
2. Athanasopoulos, G., Ahmed, R.A., Hyndman, R.J.: Optimal combination forecasts for hierarchical time series. Comput. Stat. Data Anal. **55**(9), 2579–2589 (2011)
3. Atiya, A.F.: Why does forecast combination work so well? Int. J. Forecast. **36**(1), 197–200 (2020)
4. Athanasopoulos, G., Ahmed, R.A., Hyndman, R.J.: Hierarchical forecasts for Australian domestic tourism. Int. J. Forecast. **25**(1), 146–166 (2009)
5. Barrow, D.K., Kourentzes, N.: Distributions of forecasting errors of forecast combinations: implications for inventory management. Int. J. Prod. Econ. **177**, 24–33 (2016)
6. Bates, J.M., Granger, C.W.: The combination of forecasts. J. Oper. Res. Soc. **20**(4), 451–468 (1969)
7. Box, G.E., Jenkins, G.M., Reinsel, G.: Time Series Analysis: Forecasting and Control. Holden-day, San Francisco (1970)
8. Buckland, S.T., Burnham, K.P., Augustin, N.H.: Model selection: an integral part of inference. Biometrics 603–618 (1997)
9. Chambers, J.C., Mullick, S.K., Smith, D.D.: How to choose the right forecasting technique. Harv. Bus. Rev. **49**, 45–74 (1971)
10. Chatfield, C.: Model uncertainty and forecast accuracy. J. Forecast. **15**(7), 495–508 (1996)
11. Clemen, R.T.: Combining forecasts: a review and annotated bibliography. Int. J. Forecast. **5**(4), 559–583 (1989)
12. Clements, M.P., Hendry, D.F.: Modelling methodology and forecast failure. Econ. J. **5**(2), 319–344 (2002)
13. Di Fonzo, T., Girolimetto, D.: Cross-temporal forecast reconciliation: optimal combination method and heuristic alternatives. Int. J. Forecast. (2021)
14. Fair, R.C., Shiller, R.J.: Comparing information in forecasts from econometric models. Am. Econ. Rev. **3**, 75–389 (1990)
15. Gardner, E.S., Jr.: Exponential smoothing: the state of the art-Part II. Int. J. Forecast. **22**(4), 637–666 (2006)

16. Dielman, T.E.: Least absolute value regression: recent contributions. J. Stat. Comput. Simul. **75**(4), 263–286 (2005)
17. Granger, C.W., Joyeux, R.: An introduction to long-memory time series models and fractional differencing. J. Time Ser. Anal. **1**(1), 15–29 (1980)
18. Granger, C.W.: Long memory relationships and the aggregation of dynamic models. J. Econ. **14**(2), 227–238 (1980)
19. Granger, C.W., Ramanathan, R.: Improved methods of combining forecasts. J. Forecast. **3**(2), 197–204 (1984)
20. Hsiao, C., Wan, S.K.: Is there an optimal forecast combination? J. Econ. **178**, 294–309 (2014)
21. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L.: Optimal combination forecasts for hierarchical time series. Comput. Stat. Data Anal. **55**(9), 2579–2589 (2011)
22. Hyndman, R.J., Lee, A.J., Wang, E.: Fast computation of reconciled forecasts for hierarchical and grouped time series. Comput. Stat. Data Anal. **97**, 16–32 (2016)
23. Hubrich, K.: Forecasting euro area inflation: does aggregating forecasts by HICP component improve forecast accuracy? Int. J. Forecast. **21**(1), 119–136 (2005)
24. Schwarzkopf, A.B., Tersine, R.J., Morris, J.S.: Top-down and bottom-up forecasting in S&OP. Int. J. Forecast. J. Bus. Forecast. **25**(2), 14–16 (2006)
25. Makridakis, S.: Why combining works? Int. J. Forecast. **5**(4), 601–603 (1989)
26. Mancuso, A.C.B., Werner, L.: Review of combining forecasts approaches. Indep. J. Manag. Prod. **4**(1), 248–277 (2013)
27. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 competition: 100,000 time series and 61 forecasting methods. Int. J. Forecast. **36**(1), 54–74 (2020)
28. Makridakis, S., Hibon, M.: The M3-competition: results, conclusions and implications. Int. J. Forecast. **16**(4), 451–476 (2000)
29. Marcellino, M., Stock, J.H., Watson, M.W.: Macroeconomic forecasting in the euro area: country specific versus area-wide information. Eur. Econ. Rev. **47**(1), 1–18 (2003)
30. Newbold, P., Granger, C.W.: Experience with forecasting univariate time series and the combination of forecasts. J. R. Stat. Soc.: Ser. (Gen.) **137**(2), 131–146 (1974)
31. Palm, F.C., Zellner, A.: To combine or not to combine? Issues of combining forecasts. J. Forecast. **11**(8), 687–701 (2016)
32. Sessions, D.N., Chatterjee, S.: The combining of forecasts using recursive techniques with non-stationary weights. J. Forecast. **8**(3), 239–251 (1989)
33. Schwarzkopf, A.B., Tersine, R.J., Morris, J.S.: Top-down versus bottom-up forecasting strategies. Int. J. Forecast. Int. J. Prod. Res. **26**(11), 1833–1843 (1988)
34. Spiliotis, E., Petropoulos, F., Assimakopoulos, V.: Improving the forecasting performance of temporal hierarchies. PLoS One **14**(10) (2019)
35. Stock, J.H., Watson, M.W.: A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Hengle, R.F., White, H. (eds.), Festschrift in honor of Clive Granger, pp. 1–44 (2001)
36. Stock, J.H., Watson, M.W.: Combination forecasts of output growth in a seven-country data set. Int. J. Forecast. **23**(6), 405–430 (2004)
37. Taieb, S.B., Taylor, J.W., Hyndman, R.J.: Coherent probabilistic forecasts for hierarchical time series. In: International Conference on Machine Learning, pp. 3348–3357 (2017)
38. Van Erven, T., Cugliari, J.: Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts (2015)
39. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J.: Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J. Am. Stat. Assoc. **114**(526), 804–819 (2019)
40. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J.: Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J. Am. Stat. Assoc. **114**(526), 804–819 (2019)

Frequency Domain Clustering: An Application to Time Series with Time-Varying Parameters

Raffaele Mattera and Germana Scepi

Abstract Time series distribution parameters, such as mean and variance, are usually used as features for clustering. In this paper, starting from the hypothesis that the distributional features of the time series are time-varying, a frequency domain clustering approach based on time-varying parameters is applied. Under a specified probability distribution, we estimate the time-varying parameters with the Generalized Autoregressive Score (GAS) model and cluster time series data according to a distance based on the obtained parameters' frequency domain representation. Previous studies showed that frequency domain approaches are particularly useful for clustering financial time series data. Considering both simulated and real time series data, we compare the performances of the frequency domain clustering on time-varying parameters with those obtained with time-domain benchmark procedures.

1 Introduction

Clustering is one of the most common approaches used to highlight similar patterns in a given dataset. Clustering of time series has much relevant application in different domains of sciences, such as medicine [8, 37], engineering [38], economics [1] or finance [19, 29]. The primary purpose of clustering is to group statistical units according to a similarity measure or a distance. The main problem in time series clustering is the computation of a proper distance across time series [20].

The different developed clustering approaches can be classified into three main categories [23]: observation-based, feature-based and model-based. The first class of approaches uses raw data by computing the distances directly on the basis of the observed time series. However, previous studies share that a correct classification can be achieved by considering several time series characteristics [36]. The feature-

R. Mattera (✉)
Department of Social and Economic Sciences, Sapienza University of Rome, Rome, Italy
e-mail: raffaele.mattera@uniroma1.it

G. Scepi
Department of Economics and Statistics, University of Naples "Federico II", Naples, Italy
e-mail: scepi@unina.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_15

based approaches aim to group time series by accounting for time series' features such as the autocorrelation function (ACF) [12], the periodogram [5, 22] or the cepstral coefficients [11, 33]. Similarly, the trend, the seasonality or the time series' distribution characteristics—such as mean, variance, skewness and kurtosis—can be considered interesting features for clustering [17, 21, 25]. The model-basedmodel-based clustering assumes that the time series are generated by the same statistical model but differ for the estimated parameters. In this point of view, the underlying models can be the ARIMA [28] and the GARCH [14, 26], in the univariate setting, or the Dynamic Conditional Correlation (DCC) [27], in the case of multivariate setting. Another model-based approach assumes that the same probability distribution generates the time series but with different parameters [15, 24]. However, we have to note that feature and model-based approaches are closely related because the parameters estimated by statistical models can be seen as features that fully characterize the time series [2].

In most previous studies, the model parameters are assumed to be static. However, this assumption doesn't seem to hold in the time series framework [7]. To the best of our knowledge, the first paper that, after the specification of an underlying probability density function, cluster time series according to a target time-varying parameter is [7]. A Euclidean distance based on the time-varying parameter's auto-correlation structure was used to measure the dissimilarity among time series. However, [5] demonstrated that the frequency domain distance based on the periodogram ordinates provides a better clustering quality than ACF-based metrics. Moreover, an extensive study carried out by [13] demonstrated that the frequency domain distances are the most effective for clustering generic time series.

This paper applies a model-based clustering approach based on a frequency domain distance among time-varying parameters. Following [7], by assuming a Gaussian probability density function, we use the Generalized Autoregressive Score (GAS) [10] for modelling the dynamics of the time-varying parameters and for their estimations. The GAS is a very general statistical model that considers the score function of the predictive model density as the driving mechanism for time-varying parameters. A broad class of GARCH-type processes are special cases of the GAS [10]. Since the GAS is based on the score, it exploits the data's complete density structure rather than just a few moments.

We estimate the time-varying parameters with the GAS model through in-sample predictions [6, 7] and compute a dissimilarity measure based on the estimated time-varying parameter's periodogram. We apply the Partition Around Medoids (PAM) algorithm for clustering according to this distance.

To show the performances of the frequency domain clustering on time-varying parameters, we compare the results obtained on real and simulation data, considering ACF-based procedures as a benchmark.

The paper's structure is the following: in the next section, the clustering procedure is presented, while in Sect. 3, a simulation study is provided. Section 4 shows an application to real financial time series, discussing the applicability of the so obtained groups for portfolio selection. Section 5 shows some final remarks.

2 The Clustering Procedure

In what follows, a Gaussian predictive density is assumed. The Gaussian distribution well describes many real time series data. However, several alternatives to the Gaussian model are often considered for financial time series [4]. Nevertheless, when the time series are observed at low frequencies (e.g. monthly or quarterly), the stock returns show a Gaussian density. This *stylized fact* is called *aggregate Gaussianity* [9]. Therefore, we consider that the Gaussian assumption can be appropriate for modelling financial time series observed at low frequencies.

Assuming that the time series follow a Gaussian density, the clustering procedure, adopted in this paper, is based on two steps. The first one consists of the estimation of the time-varying parameters μ_t and σ_t^2, by means of the Generalized Autoregressive Score (GAS) model [10]. In the second step, we apply the Partition Around Medoid (PAM) algorithm for clustering time series with respect to the chosen target parameter. As discussed in the introduction, we adopt a frequency domain approach in the definition of the distance.

2.1 The GAS Model with Gaussian Density

Let $\mathbf{Y}$ be the matrix containing $N(n = 1, \ldots, N)$ y_t time series on the columns and $T(t = 1, \ldots, T)$ time in the rows. Let now suppose that the following Gaussian density generates each n-th time series in the sample with different parameters μ_t and σ_t^2:

$$p(y_t; \mu_t, \sigma_t^2) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-(y_t - \mu_t)^2 / 2\sigma_t^2} \tag{1}$$

where μ_t is the time-varying mean, and σ_t^2 is the time-varying variance. We propose to cluster time series according to Gaussian time-varying parameters.

The GAS model's information set at a given point in time t, $\mathcal{F}_t$, is obtained by the previous realizations of the time series y_t and the time varying parameters $f_t = (\mu_t, \sigma_t^2)$. The Guassian-GAS of order one can be written as:

$$f_t = \omega + \mathbf{A} s_{t-1} + \mathbf{B} f_{t-1} \tag{2}$$

where:

$$f_t = \begin{pmatrix} \mu_t \\ \sigma_t^2 \end{pmatrix}, \quad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

In particular, ω is real vector with constants and $\mathbf{A}$ and $\mathbf{B}$ are diagonal matrices.

Moreover, s_t is the *scaled* score vector of the conditional density in a time t. The scaled score s_t is defined as:

$$s_t = S_t \cdot \nabla_t \tag{3}$$

where S_t is a positive definite scaling matrix known at time t and ∇_t is the score of the time series:

$$\nabla_t = \frac{\partial \log p(y_t | f_t, \mathcal{F}_t; \theta)}{\partial f_t} \tag{4}$$

Usually S_t is set to be equal to the Identiy matrix or to the inverse of Fisher information matrix [10]. Following [7], in the case of Gaussian density (1) we have that the conditional score vectors is given by:

$$\nabla_t^{(\mu)} = \frac{(y_t - \mu_t)}{\sigma_t^2}$$

$$\nabla_t^{(\sigma)} = \frac{(y_t - \mu_t)^2}{2\sigma_t^4} - \frac{T}{2\sigma_t^2}$$

A useful property of the GAS model is that the vector of parameters θ can be estimated by maximum likelihood approach because the density (1) is specified a priori. Once the parameters are estimated, we can reconstruct the time pattern of the parameters $f_t = \left(\mu_t, \sigma_t^2\right)$ by means of in-sample predictions [6, 7].

2.2 *Frequency Domain Clustering*

Following a simple raw-data based approach, we could consider a classical Euclidean distance among the time series of the time-varying parameters. However, this distance is not appropriate when dealing with time series because it does not consider essential features. For this reason, [7] used an autocorrelation-based Euclidean dissimilarity for clustering time series based on the time-varying parameters.

However, [5, 13] demonstrated that the frequency domain approaches are better suited for clustering time series than ACF-based metrics. Moreover, previous studies [e.g. see 4, 11] highlighted the particularly good results of frequency domain approaches for clustering peculiar time series, like financial ones. Therefore, in the following, we adopt a frequency domain approach for clustering time series with time-varying parameters.

We consider the *spectral density function* (also called the *spectrum*) as the frequency domain representation of the selected time-varying parameter [22, 35]. An

unbiased estimate of the spectral density is given by the so-called periodogram [16]. Let:

$$\mathrm{P}(\lambda)_{n,j} = \frac{1}{T}\left|\sum_{t=1}^{T} f_{n,j,t} e^{-i\lambda t}\right|^2 \quad \lambda \in [0, \pi] \tag{5}$$

be the periodogram of a given estimated conditional j-th moment of the n-th time series at frequencies $\lambda = 2\pi l/T$, given $\{l = 1, \ldots, T/2\}$. Therefore, a first frequency domain dissimilarity is based on the following Euclidean distance between periodiogram ordinates:

$$d_{n,n'} = \left|\left|\mathrm{P}(\lambda)_{n,j} - \mathrm{P}(\lambda)_{n,j'}\right|\right| \tag{6}$$

As argued by [5], another alternative distance measure based on the periodiogram is the following:

$$d_{n,n'} = \left|\left|\mathrm{NP}(\lambda)_{n,j} - \mathrm{NP}(\lambda)_{n,j'}\right|\right| \tag{7}$$

where $\mathrm{NP}(\lambda)_{n,j} = \mathrm{P}(\lambda)_{n,j}/\sigma^2$ is the normalized periodogram. The normalization induced by the time series static variance σ^2 can be introduced if we are not interested in the time series scales but only in their correlation structure.

Once the dissimilarity matrix is defined, several algorithms can be used for clustering. The most common is the k-means algorithm. The k-means relies on an iterative scheme based on the minimization of the following objective function:

$$\min : \sum_{i=1}^{N}\sum_{c=1}^{C} d_{i,c} \tag{8}$$

where N is the number of the time series to be clustered, C is the number of clusters (a priori fixed), c represents the centre such that $d_{i,c}$ is the distance between each time series i from the centroid of the c-th cluster.

The c-th centroid time series is a fictitious object computed as the mean of the other time series belonging to the c-th cluster. However, using a fictitious centroid reduces the interpretability of the clusters. The *Partitioning Around Medoid* (PAM) algorithm, instead, considers a real-time series as a cluster's medoid. Together with better interpretability of the results, the PAM algorithm is also better in terms of execution time [31]. Therefore, we adopt a clustering approach based on the PAM algorithm.

The main drawback of the PAM algorithm is the a priori selection of the number of clusters C. Following previous studies, we consider the Average Silhouette Width (ASW) criterion [32]:

$$\text{ASW} = \frac{1}{N}\sum_{n=1}^{N} S_n$$

where:

$$S_n = \frac{(b_n - a_n)}{\max\{b_n, a_n\}} \tag{9}$$

The value a_n is the average distance of the n-th unit to the other elements belonging to its cluster, while b_n is the average distance of the n-th time series to the nearest cluster to which it is not assigned. A large value of S_n means that b_n is much larger than a_n. Consequently, the n-th observation is much closer to the one in its cluster than to the neighbouring one. Hence, large values of S_n indicate a good clustering quality. In this sense, an optimal clustering maximizes the Silhouette [3].

The adopted clustering approach uses a target time-varying parameter for clustering time series. Therefore, we can obtain different classifications according to the different distribution parameters. Based on the specific applications, the researchers can potentially classify dynamic objects with similar time-varying means or variances.

3 Simulation Study

In what follows, we compare the classification accuracy of the time-varying parameters based clustering with two classical clustering techniques on raw time series [5, 12]. Moreover, we compare the frequency domain approach with the one, proposed by [7], that employs an ACF-based distance.

For explaining how the benchmark time-domain models are defined, we provide some preliminaries on the ACF-based distance. Let define $\hat{\rho}_{l,n}$ the estimated autocorrelation at lag l of a given n-th time series ($n = 1, \ldots, N; t = 1, \ldots, T$). The estimated autocorrelation of a given time series $y_{n,t}$, called $\hat{\rho}_{l,n}$, can be obtained with the usual estimator:

$$\hat{\rho}_{l,n} = \frac{\sum_{t=l+1}^{T} \left(y_{n,t} - \bar{y}_n\right)\left(y_{n,t-l} - \bar{y}_n\right)}{\sum_{t=1}^{T} \left(y_{n,t} - \bar{y}_n\right)^2} \tag{10}$$

where $\bar{y}_n$ is the mean of the n-th time series over T. Given a pair of time series $y_{n,t}$ and $y_{n\prime,t}$, the ACF-based Euclidean distance can be defined as:

$$d_{n,n\prime} = \left\| \hat{\rho}_{n,l} - \hat{\rho}_{n\prime,l} \right\| \tag{11}$$

3e choose as benchmarks: (a) the ACF-based clustering approach proposed by [12], which considers the distance (11); (b) the ACF-based clustering approach proposed by [7], that calculates the auto-correlation (10) among time varying parameters and (c) the frequency domain-based clustering approach proposed by [5], that calculates a periodogram-based distance on the raw time series.

Since in this case the ground truth is available, we measure the quality of classification of the different clustering approaches by means of the Rand index [30]. Let $\mathbf{Y}$ the matrix of dimension $T \times N$ of the N time series, a clustering $\mathbf{K}$ on $\mathbf{Y}$ allows to partition the set of time series into $C (i = 1, \ldots, C)$ non-overlapping groups $\{K_1, K_2, \ldots, K_i, \ldots, K_C\}$, where $\cup_{i=1}^{C} K_i = Y$ and $K_i \cap K_{i\prime} = \emptyset$ for $i \neq i\prime$. Let us consider an alternative partition $\tilde{\mathbf{K}}$. We define N_{11} the number of objects that are in the same cluster both in $\mathbf{K}$ and $\tilde{\mathbf{K}}$, N_{00} those that are in different clusters both in $\mathbf{K}$ and $\tilde{\mathbf{K}}$, N_{01} the ones that are in the same cluster for $\mathbf{K}$ but in different clusters for $\tilde{\mathbf{K}}$, and N_{10} the number of time series that are in different clusters in $\mathbf{K}$ but in the same cluster in $\tilde{\mathbf{K}}$.

N_{11} and N_{00} can be used as measures of the degree of agreement between $\mathbf{K}$ and $\tilde{\mathbf{K}}$ while, conversely, N_{01} and N_{10} can be seen as measures of disagreement. The Rand index [30] is defined as:

$$\text{RI} = \frac{(N_{00} + N_{11})}{\binom{N}{2}} \tag{12}$$

However, since the RI often lies within the range of [0.5, 1], [18] proposed the following adjustment:

$$\text{ARI} = \frac{2\,(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})\,(N_{01} + N_{11}) + (N_{00} + N_{10})\,(N_{10} + N_{11})} \tag{13}$$

that is always between the range [0, 1]. It is called the Adjusted Rand Index (ARI). An ARI value close to 0 indicates randomness in the partition, while a value close to 1 indicates a good classification. The comparison is made in terms of the average Adjusted Rand Index (ARI) over 300 trials as in [13].

In this section the comparison is conducted on simulated data. The simulation scenario is generated as follows. We simulate $N = 5$ time series that are Normally distributed with a time-varying mean $\mu_{1,t}$ and variance $\sigma^2_{1,t}$ whose process is given by the GAS with parameters:

$$\omega_1 = (0.0490, 0.0154); \quad \mathbf{A}_1 = \begin{pmatrix} 0.0001 & 0 \\ 0 & 0.0534 \end{pmatrix}; \quad \mathbf{B}_1 = \begin{pmatrix} 0.0485 & 0 \\ 0 & 0.9891 \end{pmatrix}$$

Table 1 Clustering results: average Adjusted Rand Index

Clustering approach	$T = 50$	$T = 250$	$T = 1000$	$T = 2500$
Frequency domain approaches with time-varying parameters				
Periodogram-based distance on μ_t targeting parameter	**0.7152**	**0.6728**	**0.6934**	0.7193
Periodogram-based distance on σ_t^2 targeting parameter	0.0793	0.0790	0.0657	0.0483
Benchmarks				
ACF-based distance on μ_t targeting parameter [7]	0.0137	0.0053	0.0082	0.0101
ACF-based distance on σ_t^2 targeting parameter [7]	0.0200	0.0962	0.5544	**0.8968**
ACF-based distance on raw time series y_t [12]	0.0107	0.0027	0.0008	0.0011
Periodogram-based distance on raw time series y_t [5]	0.0005	0.0006	0.0002	0.0000

Note Table reports the average Adjusted Rand Index (ARI). The best approach is highlighted with the bold font

Then, we simulate another set of $N = 5$ Gaussian time series with $\mu_{2,t}$ and $\sigma_{2,t}^2$ generated by a GAS process with the following parameters:

$$\omega_2 = (0.0840, 0.0456); \quad \mathbf{A}_2 = \begin{pmatrix} 0.00001 & 0 \\ 0 & 0.0139 \end{pmatrix}; \quad \mathbf{B}_2 = \begin{pmatrix} 0.0660 & 0 \\ 0 & 0.0968 \end{pmatrix}$$

These parameters are calibrated according to two different, randomly selected, real time series belonging to the Dow Jones 30 financial market index. We consider four different scenarios in terms of time series' length $T = \{50, 250, 1000, 2500\}$. The results are shown in Table 1.

Table 1 shows that the frequency domain clustering approach, where the mean is chosen as the target parameter, works much better than all the considered alternatives in a scenario with medium ($T = 250, 1000$) or short ($T = 50$) time series.

By increasing the time series length ($T = 2500$), we observe that the adopted frequency domain approach becomes less accurate than the ACF-based approach of [7] with variance as the target parameter. Nevertheless, in a scenario with a very large time series, the frequency domain approach with mean as target provides a good classification with a relatively high ARI.

Overall, the clustering approaches based on time-varying parameters provide better classification than the raw time series ones. This evidence holds for any time series length. In real applications, especially for economic and social time series that are usually sampled at low frequencies, it is difficult to handle with time series larger than $T = 1000$. This makes the frequency domain clustering approach, based on conditional mean targeting, better suited for classifying many real time series.

4 Application to Financial Time Series

We consider the monthly returns of the stocks included in the S&P500 index between 1-th January 2001 and 1-th January 2021. We excluded stocks with missing values. Hence, the number of considered time series is $N = 376$, all with length $T = 240$. The list of the considered stocks is shown in Tab. 2.
For finding groups of homogeneous stocks, we compare the results of the same clustering approaches used with simulated data.

Because of the high dimensionality of the dataset, showing the time series of the estimated time-varying parameters for all the stocks is prohibitive. Therefore, as an example, we focus on the stocks belonging to the Financial sector.

The time-varying mean series are shown in Fig. 1, while the time-varying variance series are reported in Fig. 2.

We note a high degree of heterogeneity in terms of time evolution for both mean and variance series. Some time series (e.g. AFL, AXP, CB, etc.) do not show a trend, while others are characterized by positive or negative trends, with possible structural breaks (e.g. MMC, PGR, SPGI, etc.).

Similarly, Fig. 2 shows the heterogeneity in terms of returns' variances. Most of them show an outlier, characterized by a considerable increase in variance during the 2008 financial crisis, while some stocks (e.g. CMA, BK, TRV, etc.) show a very different pattern.

We underlay that this high degree of heterogeneity holds for all the stocks in the sample.

We consider the six alternative clustering approaches and their final cluster assignment. According to the Average Silhouette Width (ASW), the number of clusters is chosen as explained in Sect. 2. Figure 3 shows the ASW for each of the alternative clustering approaches.

We set the maximum number of clusters to be $C = 11$ as the number of the sectors (see Tab. 2). All the clustering algorithms suggest the presence of $C = 2$ homogeneous groups.

Table 2 List of considered stocks

Symbol	Sector	Symbol	Sector	Symbol	Sector	Symbol	Sector
MMM	Industrials	CMI	Industrials	JNJ	Health Care	DGX	Health Care
AOS	Industrials	CVS	Health Care	JCI	Industrials	RL	Consumer Discretionary
ABT	Health Care	DHI	Consumer Discretionary	JPM	Financials	RJF	Financials
ABMD	Health Care	DHR	Health Care	JNPR	Information Technology	RTX	Industrials
ATVI	Communication Services	DRI	Consumer Discretionary	KSU	Industrials	O	Real Estate
ADBE	Information Technology	DVA	Health Care	K	Consumer Staples	REG	Real Estate
AMD	Information Technology	DE	Industrials	KEY	Financials	REGN	Health Care
AES	Utilities	XRAY	Health Care	KMB	Consumer Staples	RF	Financials
AFL	Financials	DVN	Energy	KIM	Real Estate	RSG	Industrials
A	Health Care	DISH	Communication Services	KLAC	Information Technology	RMD	Health Care
APD	Materials	DLTR	Consumer Discretionary	KR	Consumer Staples	RHI	Industrials
AKAM	Information Technology	D	Utilities	LHX	Industrials	ROK	Industrials
ALK	Industrials	DOV	Industrials	LH	Health Care	ROL	Industrials
ALB	Materials	DTE	Utilities	LRCX	Information Technology	ROP	Industrials
ARE	Real Estate	DUK	Utilities	LEG	Consumer Discretionary	ROST	Consumer Discretionary
LNT	Utilities	DRE	Real Estate	LEN	Consumer Discretionary	RCL	Consumer Discretionary
ALL	Financials	DD	Materials	LLY	Health Care	SPGI	Financials
MO	Consumer Staples	DXC	Information Technology	LNC	Financials	SBAC	Real Estate
AMZN	Consumer Discretionary	EMN	Materials	LIN	Materials	SLB	Energy
AEE	Utilities	ETN	Industrials	LMT	Industrials	SEE	Materials
AEP	Utilities	EBAY	Consumer Discretionary	L	Financials	SRE	Utilities
AXP	Financials	ECL	Materials	LOW	Consumer Discretionary	SHW	Materials
AIG	Financials	EIX	Utilities	LUMN	Communication Services	SPG	Real Estate
AMT	Real Estate	EW	Health Care	MTB	Financials	SWKS	Information Technology
ABC	Health Care	EA	Communication Services	MRO	Energy	SNA	Industrials
AME	Industrials	EMR	Industrials	MAR	Consumer Discretionary	SO	Utilities
AMGN	Health Care	ETR	Utilities	MMC	Financials	LUV	Industrials
APH	Information Technology	EOG	Energy	MLM	Materials	SWK	Industrials
ADI	Information Technology	EFX	Industrials	MAS	Industrials	SBUX	Consumer Discretionary
ANSS	Information Technology	EQIX	Real Estate	MKC	Consumer Staples	STT	Financials
AON	Financials	EQR	Real Estate	MCD	Consumer Discretionary	STE	Health Care
APA	Energy	ESS	Real Estate	MCK	Health Care	SYK	Health Care

(continued)

Table 2 (continued)

Symbol	Sector	Symbol	Sector	Symbol	Sector	Symbol	Sector
AAPL	Information Technology	EL	Consumer Staples	MDT	Health Care	SIVB	Financials
AMAT	Information Technology	RE	Financials	MRK	Health Care	SNPS	Information Technology
ADM	Consumer Staples	EVRG	Utilities	MET	Financials	SYY	Consumer Staples
AJG	Financials	ES	Utilities	MTD	Health Care	TROW	Financials
T	Communication Services	EXC	Utilities	MGM	Consumer Discretionary	TTWO	Communication Services
ATO	Utilities	EXPD	Industrials	MCHP	Information Technology	TPR	Consumer Discretionary
ADSK	Information Technology	XOM	Energy	MU	Information Technology	TGT	Consumer Discretionary
ADP	Information Technology	FFIV	Information Technology	MSFT	Information Technology	TDY	Industrials
AZO	Consumer Discretionary	FAST	Industrials	MAA	Real Estate	TFX	Health Care
AVB	Real Estate	FRT	Real Estate	MHK	Consumer Discretionary	TER	Information Technology
AVY	Materials	FDX	Industrials	TAP	Consumer Staples	TXN	Information Technology
BKR	Energy	FITB	Financials	MNST	Consumer Staples	TXT	Industrials
BLL	Materials	FE	Utilities	MCO	Financials	BK	Financials
BAC	Financials	FISV	Information Technology	MS	Financials	CLX	Consumer Staples
BAX	Health Care	FMC	Materials	MSI	Information Technology	COO	Health Care
BDX	Health Care	F	Consumer Discretionary	NTAP	Information Technology	HSY	Consumer Staples
BBY	Consumer Discretionary	BEN	Financials	NWL	Consumer Discretionary	MOS	Materials
BIO	Health Care	FCX	Materials	NEM	Materials	TRV	Financials
BIIB	Health Care	GPS	Consumer Discretionary	NEE	Utilities	DIS	Communication Services
BLK	Financials	GRMN	Consumer Discretionary	NKE	Consumer Discretionary	TMO	Health Care
BA	Industrials	IT	Information Technology	NI	Utilities	TJX	Consumer Discretionary
BKNG	Consumer Discretionary	GD	Industrials	NSC	Industrials	TSCO	Consumer Discretionary
BWA	Consumer Discretionary	GE	Industrials	NTRS	Financials	TT	Industrials
BXP	Real Estate	GIS	Consumer Staples	NOC	Industrials	TRMB	Information Technology
BSX	Health Care	GPC	Consumer Discretionary	NLOK	Information Technology	TFC	Financials
BMY	Health Care	GILD	Health Care	NOV	Energy	TYL	Information Technology
CHRW	Industrials	GL	Financials	NUE	Materials	TSN	Consumer Staples
COG	Energy	GS	Financials	NVDA	Information Technology	USB	Financials
CDNS	Information Technology	GWW	Industrials	NVR	Consumer Discretionary	UDR	Real Estate
CPB	Consumer Staples	HAL	Energy	ORLY	Consumer Discretionary	UNP	Industrials
COF	Financials	HIG	Financials	OXY	Energy	UPS	Industrials

(continued)

Table 2 (continued)

Symbol	Sector	Symbol	Sector	Symbol	Sector	Symbol	Sector
CAH	Health Care	HAS	Consumer Discretionary	ODFL	Industrials	URI	Industrials
KMX	Consumer Discretionary	PEAK	Real Estate	OMC	Communication Services	UNH	Health Care
CCL	Consumer Discretionary	HSIC	Health Care	OKE	Energy	UHS	Health Care
CAT	Industrials	HES	Energy	ORCL	Information Technology	UNM	Financials
CNP	Utilities	HFC	Energy	PCAR	Industrials	VLO	Energy
CERN	Health Care	HOLX	Health Care	PKG	Materials	VTR	Real Estate
SCHW	Financials	HD	Consumer Discretionary	PH	Industrials	VRSN	Information Technology
CVX	Energy	HON	Industrials	PAYX	Information Technology	VZ	Communication Services
CB	Financials	HRL	Consumer Staples	PENN	Consumer Discretionary	VRTX	Health Care
CHD	Consumer Staples	HST	Real Estate	PNR	Industrials	VFC	Consumer Discretionary
CI	Health Care	HPQ	Information Technology	PBCT	Financials	VTRS	Health Care
CINF	Financials	HUM	Health Care	PEP	Consumer Staples	VNO	Real Estate
CTAS	Industrials	HBAN	Financials	PKI	Health Care	VMC	Materials
CSCO	Information Technology	IEX	Industrials	PRGO	Health Care	WRB	Financials
C	Financials	IDXX	Health Care	PFE	Health Care	WBA	Consumer Staples
CTXS	Information Technology	ITW	Industrials	PNW	Utilities	WMT	Consumer Staples
CMS	Utilities	ILMN	Health Care	PXD	Energy	WM	Industrials
KO	Consumer Staples	INCY	Health Care	PNC	Financials	WAT	Health Care
CTSH	Information Technology	INTC	Information Technology	POOL	Consumer Discretionary	WEC	Utilities
CL	Consumer Staples	IBM	Information Technology	PPG	Materials	WFC	Financials
CMCSA	Communication Services	IFF	Materials	PPL	Utilities	WELL	Real Estate
CMA	Financials	IP	Materials	PG	Consumer Staples	WST	Health Care
CAG	Consumer Staples	IPG	Communication Services	PGR	Financials	WDC	Information Technology
COP	Energy	INTU	Information Technology	PLD	Real Estate	WAB	Industrials
ED	Utilities	ISRG	Health Care	PTC	Information Technology	WY	Real Estate
STZ	Consumer Staples	IVZ	Financials	PEG	Utilities	WHR	Consumer Discretionary
CPRT	Industrials	IRM	Real Estate	PSA	Real Estate	XEL	Utilities
GLW	Information Technology	JBHT	Industrials	PHM	Consumer Discretionary	XLNX	Information Technology
COST	Consumer Staples	JKHY	Information Technology	PVH	Consumer Discretionary	YUM	Consumer Discretionary
CCI	Real Estate	J	Industrials	QCOM	Information Technology	ZBRA	Information Technology
CSX	Industrials	SJM	Consumer Staples	PWR	Industrials	ZION	Financials

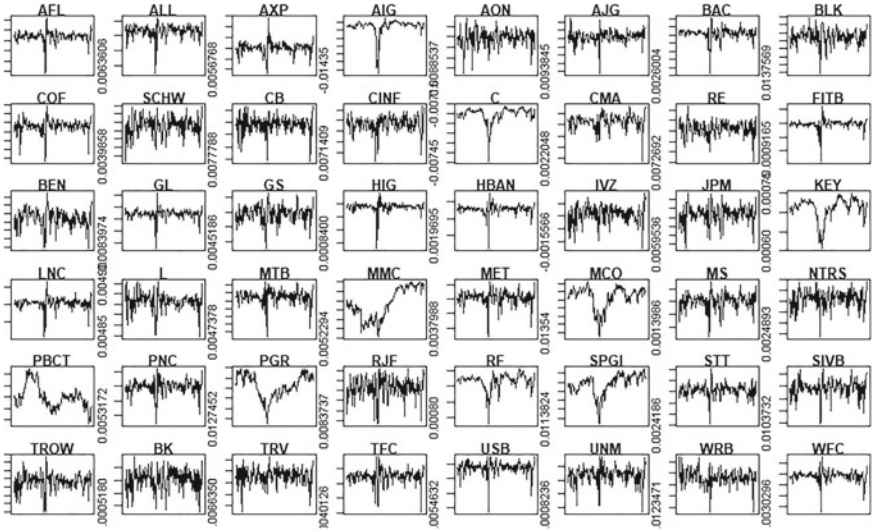


Fig. 1 Time-varying mean for Financial sector time series

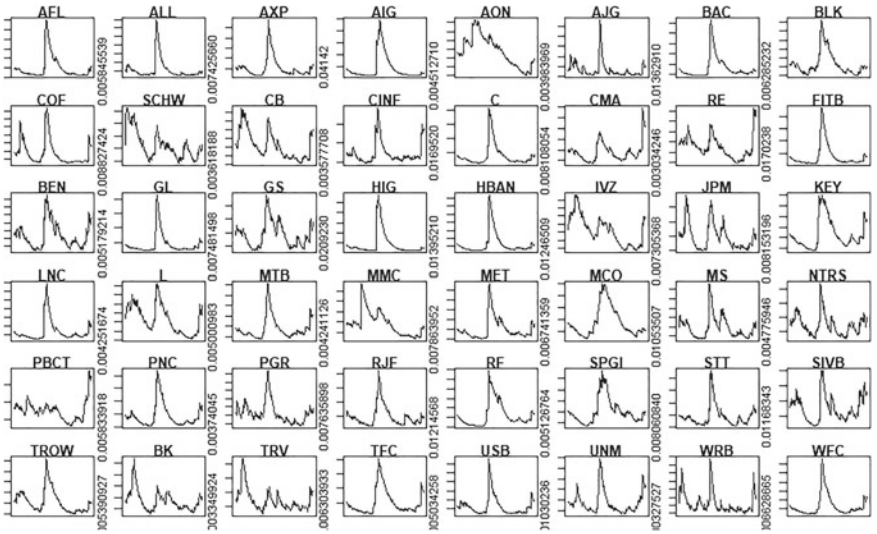


Fig. 2 Time-varying variance for Financial sector time series

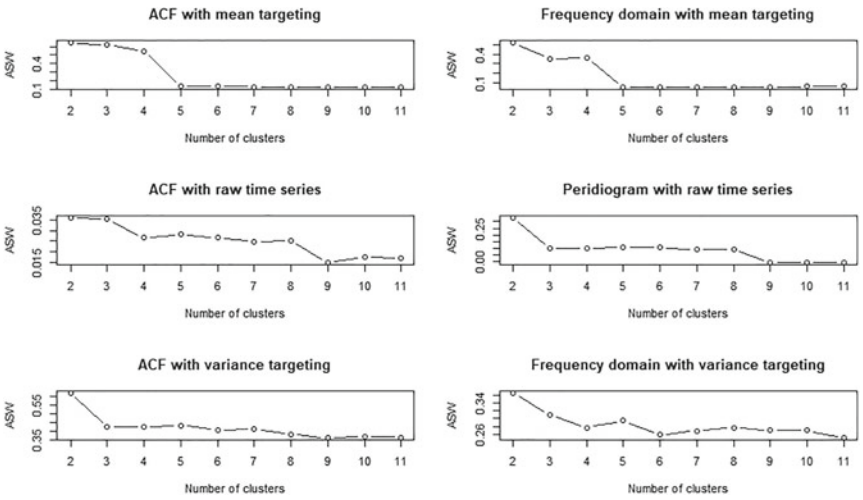


Fig. 3 ASW for the alternative clustering approaches

Moreover, Fig. 3 highlights that the best partitions are obtained by considering the approaches with time-varying parameters.

As explained previously, the ASW is an indicator of clustering quality since it measures the proximity between the units placed within the same group and the dissimilarity with those set in different clusters. High values of the ASW mean that the stocks placed in the same group are very similar and, at the same time, are very different from those belonging to other clusters.

Figure 3 shows that the best partitions are obtained with the frequency domain clustering approaches. Indeed, the ACF-based algorithms perform poorly, in terms of ASW, with respect to those based on the frequency domain features of the time series. The ASW of the first is equal to 0.035, versus the ASW of 0.25 associated with the periodogram-based distance. However, we don't note relevant differences among the chosen target parameters. Both mean and variance-based clustering approaches seem to be equally good in partitioning this dataset.

Hence, we exploit the main financial characteristics of the so obtained clusters. Table 3 shows the average Sharpe ratio of the stocks included in each group for all the considered clustering algorithms.

The Sharpe ratio [34] is a well-known measure of performance implemented by financial analysts. The Sharpe ratio of an stock i, that we call θ_i, is defined as the ratio between its average return $\bar{\mu}$ and the volatility $\bar{\sigma}$ over a given time period:

$$\theta_i = \frac{\bar{\mu}_i}{\bar{\sigma}_i}$$

Table 3 Cluster analysis: sharpe ratios

	Cluster 1		Cluster 2	
	No. Stocks	Sharpe	No. Stocks	Sharpe
Frequency domain approaches with time-varying parameters				
Periodogram-based distance on μ_t targeting parameter	325	0.107	51	0.096
Periodogram-based distance on σ_t^2 targeting parameter	179	0.108	197	0.103
Benchmarks				
ACF-based distance on μ_t targeting parameter [7]	308	0.109	68	0.089
ACF-based distance on σ_t^2 targeting parameter [7]	116	0.109	260	0.104
ACF-based distance on raw time series y_t [12]	258	0.103	118	0.109
Periodogram-based distance on raw time series y_t [5]	116	0.062	260	0.125

By comparing two stocks i and j, we say that i is more attractive if its Sharpe ratio exceeds the one of j, i.e. $\theta_i > \theta_j$.

All the clustering approaches highlight two groups of stocks characterized by similar average financial performance, but they differ in size and composition.

The cluster 2, obtained with the standard frequency domain approach of [5] has the highest Sharpe ratio while the cluster 1 has the lowest performance.

Similar performances characterize the clustering approaches with time-varying parameters in terms of Sharpe ratios. Nevertheless, the frequency domain approaches with targeting seem to have better characteristics regarding the time domain alternatives.

The frequency domain approach with mean targeting provides a higher average Sharpe ratio than the time-domain approach of [7] (10.1% vs. 9.9%). The frequency domain approach with variance targeting offers the most balanced groups in terms of size. Therefore, if a balanced group of stocks is preferred, an ideal investor would take the clustering resulting from the frequency-domain approach with variance targeting.

Another important aspect from the investor's point of view is the correlation structure of the objects placed within each cluster. Figure 4 shows the correlation

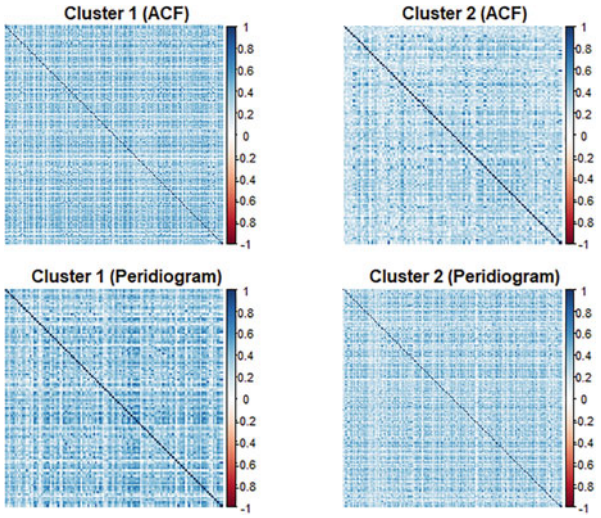


Fig. 4 Correlation structure within the clusters: ACF-based (top) and periodogram-based (bottom)

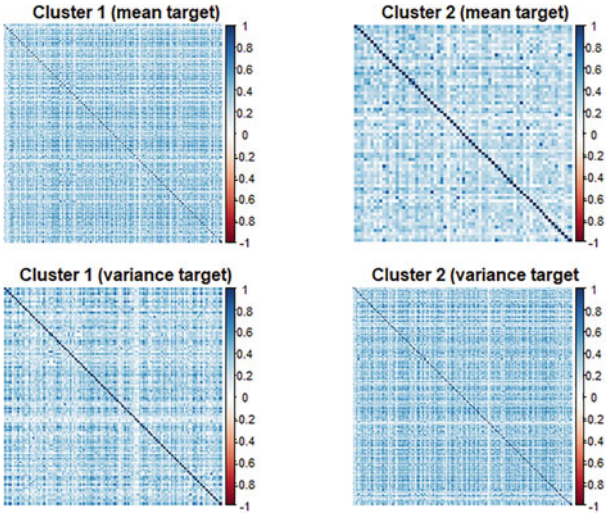


Fig. 5 Correlation structure within the clusters: mean target (top) and variance target (bottom)—time domain approach with targeting

structure for the time and frequency domain distances on raw time series, while Fig. 5 and Fig. 6 show the correlations for the parameter targeting approaches in time and frequency domain respectively.

It is well known that investors would prefer clusters with a low within-correlation structure. Dark blue colours mean positive correlations for all the figures, while

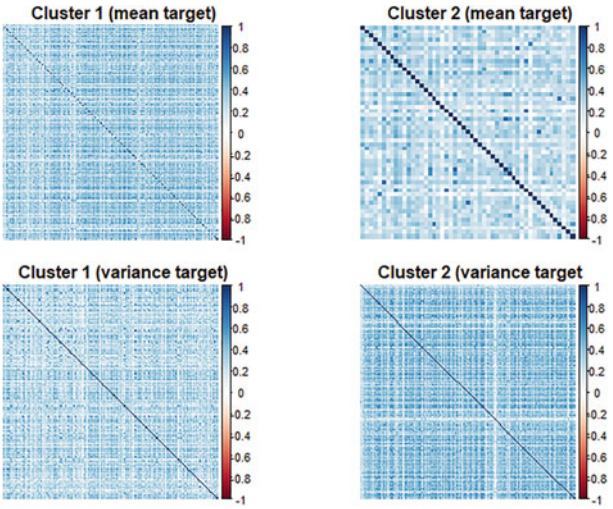


Fig. 6 Correlation structure within the clusters: mean target (top) and variance target (bottom)—frequency domain approach with targeting

lighter colours suggest weak correlations. Figure 4 shows that, for both the standard approaches, the stocks belonging to the two groups are highly correlated. Cluster 2 of the periodogram-based approach is the only exception.

By comparing the results of Fig. 4 with those of Fig. 5 and Fig. 6, we observe that the clusters in the latter two figures show more weak correlations. This evidence makes the clusters obtained with parameter targeting more attractive for investors.

Let's now compare the Fig. 5 and Fig. 6. It is evident that, for all the clusters and for both targeting parameter (either mean or variance), the colors in Fig. 6 are lighter than those in Fig. 5. This means, in other words, that the within-cluster correlation is much weaker in the case of frequency domain clustering with targeting rather than the time domain alternative.

Moreover, inspecting in more detail Fig. 6, we observe that the clusters with mean targeting are characterized by a weaker within-group correlation (see Fig. 6 on the top). Together with the clusters' size balance and Sharpe ratio, these results make this second clustering approach the most attractive for investors.

5 Final Remarks

This paper discusses an approach to clustering time series with time-varying distribution parameters, estimated through the Generalized Autoregressive Score of [10].

Since the parameters are time series themselves, the degree of similarity of their time patterns can be exploited in different ways. [7] used an autocorrelation-based

Euclidean dissimilarity for clustering time series with time-varying parameters. However, [5] demonstrated that the frequency domain distance, based on the periodogram ordinates, provides a better clustering quality than ACF-based metrics. Moreover, an extensive study carried out by [13] demonstrated that the frequency domain distances are the most effective for clustering generic time series. In this paper, we study the performances of clustering approaches based on frequency-domain features of the target time-varying parameter.

The simulation results show that the frequency domain approach based on time-varying parameters overperforms the approaches based on raw time series. Moreover, it also provides a classification more accurate than the time domain approach proposed by [7] when medium and short-sized series are involved with mean targeting. This is the case of economic and social time series, which are usually sampled at low frequencies (i.e. yearly or quarterly). The simulation data study notes that the considered benchmarks, not based on time-varying parameters, perform poorly. We think that these weak performances can be derived by the complexity of the Data Generating Process underlying the time series with time-varying parameters. Simulations conducted by previous studies [e.g. 13] mainly consider ARMA-based DGPs. Differently, we show that, if the DGP is based on time variation in the parameters, clustering approaches based on the ACF or the periodogram of the raw time series provide weak classifications. Furthermore, we show that a frequency domain approach should be preferred when the conditional mean is the target parameter, while a time domain approach should be used when the conditional variance is chosen as the target. This result could be explained by the peculiar form of the conditional variance's periodogram, characterized by peaks at specific frequencies, but future researches are needed to investigate it.

The analysis with real data is conducted considering the monthly stock returns included in the S&P500 Index. The possible selection of a target parameter can be helpful for investors showing different preferences on assets with low time-varying risk or with positive time-varying skewness. The results of the real data application provide similar evidence to those of the simulations, since the clustering approaches with time-varying parameters are associated with a better cluster quality, in terms of Silhouette, than the benchmarks ones. Moreover, the results with real data also show that the clusters obtained with the time-varying parameters frequency domain approach in the case of mean targeting have more attractive characteristics from the investors' point of view.

Future works can study the frequency domain approach's characteristics starting from different hypotheses. The development of a fuzzy extension of the frequency domain procedure can be helpful and the application to multivariate time series.

References

1. Ahlborn, M., Wortmann, M.: The core-periphery pattern of European business cycles: a fuzzy clustering approach. J. Macroecon. **55**, 12–27 (2018)

2. Bastos, J.A., Caiado, J.: On the classification of financial data with domain agnostic features. Int. J. Approx. Reason. **138**, 1–11 (2021)
3. Batool, F., Hennig, C.: Clustering with the average silhouette width. Comput. Stat. Data Anal. **158**, 107190 (2021)
4. Caiado, J., Crato, N.: Identifying common dynamic features in stock returns. Quant. Financ. **10**(7), 797–807 (2010)
5. Caiado, J., Crato, N., Peña, D.: A periodogram-based metric for time series classification. Comput. Stat. Data Anal. **50**(10), 2668–2684 (2006)
6. Cerqueti, R., D'Urso, P., De Giovanni, L., Giacalone, M., Mattera, R.: Weighted score-driven fuzzy clustering of time series with a financial application. Expert. Syst. Appl. **198**, 116752 (2022)
7. Cerqueti, R., Giacalone, M., Mattera, R.: Model-based fuzzy time series clustering of conditional higher moments. Int. J. Approx. Reason. **134**, 34–52 (2021)
8. Ceylan, R., Özbay, Y., Karlik, B.: A novel approach for classification of ecg arrhythmias: Type-2 fuzzy clustering neural network. Expert. Syst. Appl. **36**(3), 6721–6726 (2009)
9. Cont, R.: Empirical properties of asset returns: stylized facts and statistical issues (2001)
10. Creal, D., Koopman, S.J., Lucas, A.: Generalized autoregressive score models with applications. J. Appl. Econ. **28**(5), 777–795 (2013)
11. D'Urso, P., De Giovanni, L., Massari, R., D'Ecclesia, R.L., Maharaj, E.A.: Cepstral-based clustering of financial time series. Expert Syst. Appl. **161**, 113705 (2020)
12. D'Urso, P., Maharaj, E.A.: Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets Syst. **160**(24), 3565–3589 (2009)
13. Díaz, S.P., Vilar, J.A.: Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. J. Classification **27**(3), 333–362 (2010)
14. D'Urso, P., De Giovanni, L., Massari, R.: Garch-based robust clustering of time series. Fuzzy Sets Syst. **305**, 1–28 (2016)
15. D'Urso, P., Maharaj, E.A., Alonso, A.M.: Fuzzy clustering of time series using extremes. Fuzzy Sets Syst. **318**, 56–79 (2017)
16. Fan, J., Kreutzberger, E.: Automatic local smoothing for spectral density estimation. Scand. J. Stat. **25**(2), 359–369 (1998)
17. Fulcher, B.D., Jones, N.S.: Highly comparative feature-based time-series classification. IEEE Trans. Knowl. Data Eng. **26**(12), 3026–3037 (2014)
18. Hubert, L., Arabie, P.: Comparing partitions. J. Classification **2**(1), 193–218 (1985)
19. Iorio, C., Frasso, G., D'Ambrosio, A., Siciliano, R.: A p-spline based clustering approach for portfolio selection. Expert Syst. Appl. **95**, 88–103 (2018)
20. Liao, T.W.: Clustering of time series data-a survey. Pattern Recognit. **38**(11), 1857–1874 (2005)
21. Lubba, C.H., Sethi, S.S., Knaute, P., Schultz, S.R., Fulcher, B.D., Jones, N.S.: catch22: canonical time-series characteristics. Data Min. Knowl. Discov. **33**(6), 1821–1852 (2019)
22. Maharaj, E.A., D'Urso, P.: Fuzzy clustering of time series in the frequency domain. Inf. Sci. **181**(7), 1187–1211 (2011)
23. Maharaj, E.A., D'Urso, P., Caiado, J.: Time series clustering and classification. CRC Press (2019)
24. Mattera, R., Giacalone, M., Gibert, K.: Distribution-based entropy weighting clustering of skewed and heavy tailed time series. Symmetry **13**(6), 959 (2021)
25. Nanopoulos, A., Alcock, R., Manolopoulos, Y.: Feature-based classification of time-series data. Int. J. Comput. Res. **10**(3), 49–61 (2001)
26. Otranto, E.: Clustering heteroskedastic time series by model-based procedures. Comput. Stat. Data Anal. **52**(10), 4685–4698 (2008)
27. Otranto, E.: Identifying financial time series with similar dynamic conditional correlation. Comput. Stat. Data Anal. **54**(1), 1–15 (2010)
28. Piccolo, D.: A distance measure for classifying arima models. J. Time Ser. Anal. **11**(2), 153–164 (1990)
29. Raffinot, T.: Hierarchical clustering-based asset allocation. J. Portf. Manag. **44**(2), 89–99 (2017)

30. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
31. Rdusseeun, L., Kaufman, P.: Clustering by means of medoids. In: Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, Switzerland, pp. 405–416 (1987)
32. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
33. Savvides, A., Promponas, V.J., Fokianos, K.: Clustering of biological time series by cepstral coefficients based distances. Pattern Recognit. **41**(7), 2398–2412 (2008)
34. Sharpe, W.F.: The sharpe ratio. J. Portf. Manag. **21**(1), 49–58 (1994)
35. Vilar, J.M., Vilar, J.A., Pértega, S.: Classifying time series data: a nonparametric approach. J. Classification **26**(1), 3–28 (2009)
36. Wang, X., Smith, K., Hyndman, R.: Characteristic-based clustering for time series data. Data Min. Knowl. Discov. **13**(3), 335–364 (2006)
37. Yeh, Y.-C., Chiou, C.W., Lin, H.-J.: Analyzing ecg for cardiac arrhythmia using cluster analysis. Expert. Syst. Appl. **39**(1), 1000–1010 (2012)
38. Zhang, Z., Lai, X., Wu, M., Chen, L., Lu, C., Du, S.: Fault diagnosis based on feature clustering of time series data for loss and kick of drilling process. J. Process Control **102**, 24–33 (2021)

Heterogeneous Income Dynamics: Unemployment Consequences in Germany and the US

Raffaele Grotti

Abstract This paper studies income trajectories after unemployment and their stratification by levels of education in Germany and the United States. The literature recognizes that unemployment hits individuals' incomes differently according to the society's specific mobility regime. Accordingly, this paper investigates how the labor market, the welfare state as well as the household buffer income losses following unemployment, and how this varies across educational levels, between sexes and in comparative perspective. Empirical analyses are based on CNEF (SOEP and PSID) data and employ distributed fixed-effects models. Results show that institutions play a considerable role in reducing the consequences of unemployment. The role of each institution, however, varies across levels of education, being stronger either for the lower educated or for the higher educated individuals. These complex processes result in very similar trajectories in disposable household income across levels of education, with the exception of German men for which inequality penalizes the lower educated. For German men, while the equalizing role of the state almost perfectly counterbalances the un-equalizing role of the household, initial educational inequality in the labour market translates into inequality in final welfare as captured by disposable household income.

Keywords Income dynamics · Unemployment · Household · Distributed fixed-effects · Comparative

1 Introduction

This paper studies the unemployment consequences for individual income trajectories in Germany and the United States. In addition, it studies how welfare institutions shape income trajectories in the two countries.[1] This paper builds on DiPrete and McManus [14] who have already investigated this topic in these two countries, and

[1] In the following, I use 'unemployment' and 'job loss' interchangeably.

R. Grotti (✉)
Department of Sociology and Social Research, University of Trento, Via Verdi 26, Trento, Italy
e-mail: raffaele.grotti@unitn.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_16

especially on Ehlert [19, 20] who in addition has investigated job loss consequences across economic strata defined by household income quantiles. The current paper aims to expand existing research by studying how income trajectories following unemployment vary between individuals with different levels of education, a crucial characteristic for social stratification and stratification research.

Unemployment is likely to trigger short- and medium-term changes in income, since income from employment represents the main source of individual and household incomes [31, 32]. The extent to which job loss affects income trajectories may be influenced by many factors, at the micro, meso and macro level [12, 25, 38]. On the one hand, individuals differ significantly in their abilities to cope with critical life-course events. I therefore examine the impact of job loss comparing individuals with different levels of education.

On the other hand, the consequences of job loss are likely to vary according to the context in which individuals are embedded. In this respect, individuals are embedded in a meso level environment—the household—and in a macro context—the country. These contexts, characterized by their specific structures of opportunities and constraints, not only contribute to defining the risk of experiencing an event, but also the individual vulnerability following the event.

This paper builds on the idea that *events* interact with *attributes* in *context*. The experience of a critical life event may trigger processes of increasing inequality over the life-course if its consequences depend on the individual's position within the stratification system i.e. larger negative consequences for the least well-off [11, 13, 39]. Integrating the life-course and social stratification perspectives, the following research questions emerge:

Do income trajectories after job loss differ across educational levels? Are the negative consequences of unemployment more pronounced among already disadvantaged groups?

The interaction between events and social stratification, moreover, takes place embedded in context. Differences between countries may affect (1) the labour income trajectories coming with job loss; (2) the capacity that the household has to buffer income losses; and (3) the extent to which the welfare state buffers the losses. This raises further research questions:

To what extent do family and state shape income trajectories? Do their roles differ across educational levels? How do they operate in different contexts? Do their roles in shaping inequality between educational levels vary in different countries?

This paper compares Germany and the United States. These two countries are useful cases because they are characterized by different institutional settings [21, 22, 30], thus providing the necessary variation in market, household, and state. Yet, the countries differ in many other ways. It is therefore not possible to attribute potential differences in the outcomes to specific institutional arrangements, but only to the general macro context. Studying Germany and the US also enables the comparison of my results with those of previous studies [14].

While I build on existing studies, I go beyond them considering an unexplored dimension of social stratification, namely the level of education. As compared to a measure of stratification based on equivalent household income at the time of the

event [20], the choice of the level of education presents several advantages. First, education has the advantage of providing more analytical clarity by being theoretically more closely linked to the stratification of individual counter-mobility strategies (i.e. reemployment) across occupational and internal labor markets. This is true for two reasons: First, hypotheses about reemployment and its variation between countries are based on individual human capital; second, an individual- rather than a family-level measure (such as household income) is better suited to capture individual reemployment chances. A second advantage of using education is that education better captures patterns of assortative mating and thus is better suited for the formulation of expectations about the household buffer. However, education might come with the disadvantage of being less closely associated to the stratification of the welfare state buffer because social assistance benefits tend to be linked to household economic resources, i.e. household income.

In the following section, I discuss how the consequences of job loss may be shaped by different institutions among the two countries. Then, I will present the data and methods used. The section that follows presents the results while the last section discusses the results and draws the conclusions.

2 Job Loss and the Strategies to Buffer Its Consequences

Job loss falls under the umbrella of 'trigger events': Life-course events that may have strong implications for economic wellbeing as well as for intra-generational mobility processes, and are thus viewed as mechanisms that have the potential to (re)produce social stratification [12].

Existing literature agrees that unemployment has substantial negative consequences on earnings at the time the job is lost. At the same time, literature demonstrates that job loss consequences on workers' careers are persistent and uses the term 'scarring effect' to indicate the loss of earnings that reemployed individuals typically experience in their new job compared with their earnings before job loss. Among others, a well-known study on the United States found a persistent scarring effect of unemployment: Up to four years after losing their job workers had not completely recovered their initial earnings [42]. The same conclusion has been reached more recently [5]. Studies on Germany also report a scarring effect of job loss, although with a lower magnitude [8, 26].

2.1 Individual Counter-Mobility Strategy

The main mechanism or strategy for compensating or buffering income losses after unemployment is certainly reemployment. The 'buffering capacity' of reemployment may vary according to the time spent out of employment and, partly associated with it, the re-entry wage. The extent to which reemployment erases previous income losses,

however, is also determined by the labour market structure. Micro-level mechanisms are thus moderated by macro-level institutions and circumstances.

In Germany, the labour market is mainly segmented by occupations and jobs are identified by the content and skills that the job requires. Workers holding such skills have access to the same type of job in many firms. Literature [23, 37] refers to this characterization of labour market as occupational labour markets or OLMs. Accordingly, the educational system furnishes the (future) workers with standardized and reliable vocational qualifications and transferable skills that may be used in many firms. Labour mobility is possible between firms within occupations. Moreover, the standardized and reliable character of workers' educational qualifications should favour a good job-skills match, implying re-entry wages similar to the wages before job loss.

By contrast in the United States, the labour market is mainly segmented by firms (internal labour markets or ILMs), and skills are mainly acquired within the firm via on-the-job training.[2] In line with this, the educational system provides workers with general qualifications rather than vocational skills [15]. Therefore, in ILMs skills are much more difficult to transfer because there may not be corresponding jobs in other firms or because the access to such jobs is closed by institutional rules. Labour mobility is in this case possible between jobs within the same firm [23, 37]. Because of the firm-specific character of the workers' skills, job-skills match in ILMs should be more difficult and wages in the new job lower than wages before job loss.

Such a characterization of the two labour markets has different consequences for individuals that possess different levels of education. In OLMs, a major cleavage exists between those with and without (relevant) skills, which puts the low-educated/skilled in a particularly disadvantaged position. High-educated workers will thus receive job offers at a higher rate and better-paid jobs compared to low-educated workers: This means faster re-entry and higher wages in the new job. On the contrary, given the job-specific character of skills in ILMs, skills are not easily transferable and employers may give relatively less weight to workers' skills levels.

Therefore, considering the valuable character of skills in OLMs, in Germany I expect lower income losses in the years after job loss for the most educated individuals compared to the least educated. This is because high-educated people should experience the highest chances of reemployment and the highest wages. On the contrary, considering ILMs characteristics, in the United States this stratified pattern should not be observed to the same extent as in Germany (*Hypothesis 1*).

[2] Germany is in general considered to be characterized by OLMs because labor market segmentation is mainly based on occupation and the US by ILMs because segmentation is mainly based on firms. However, as Marsden [37] recognizes, OLMs and ILMs often coexist within the same economy, for example in different sectors. This means that ILMs can also be found in Germany and OLMs in the US.

2.2 Household Buffer

A second mechanism that can buffer the negative consequences of job loss operates at the household level. This mechanism, i.e. income pooling, refers to the availability of additional incomes from other household members, mainly the partner, which may alleviate income losses due to individual critical life events [6, 7].[3]

While income pooling should buffer income losses in general, the extent to which it works may vary between men and women, across countries and across social strata. Full-time employment of men is basically the norm in most industrialized countries. Conversely, women's labour market participation and 'employment intensity' vary across countries. In the United States, women's labour market participation has increased rapidly and we register a higher share of dual-earner couples than in Germany. Employment intensity is also higher because women are more likely to work full-time, compared to Germany where part-time employment is rather widespread [10, 17]. In Germany, women are more often secondary-earners contributing to a lower extent to the household budget [9].

This has implications for the role of the household buffer between sexes and countries. Given the strong labour market attachment of men, if the woman loses the job, her partner can considerably compensate for her income loss and smooth the economic consequences of the event. On the contrary, if the man loses the job, the women lower labour market attachment implies that she has a reduced capacity to compensate for his income lost. This should be especially true for Germany where female labour market participation and intensity are lower than in the United States. A further element that goes in the direction to expect a larger buffer for women is the gender-pay gap characterizing both countries.

Therefore, I expect the household buffer to be more effective for women than for men (*Hypothesis 2a*). Moreover, in cases where it is the man who loses the job, I expect a lower household buffer for German men compared to American men (*Hypothesis 2b*).

Household buffering capacity can also vary across levels of education. In this respect, who marries whom is crucial and Germany and the US are characterized by high and similar levels of homogamy, namely individuals are likely to form a household with a partner that shares some (socially and economically relevant) traits [3, 4, 28, 29, 33, 43]. Education is probably the most important trait in this respect.

Accordingly, given educational homogamy, I expect that the capacity of the household to compensate for income losses via partner's income increases with education in both countries (*Hypothesis 3*). In addition, low educated individuals are expected to be further disadvantaged because they are less likely to live with a partner and

[3] Although the partner represents the main potential provider of income, in the case of couple households, his or her income is not the only source of income that can contribute to the household budget. Other sources include incomes of other household members, private transfers, and private retirement incomes of retired members living in the household. Therefore, also single-headed households might benefit from the household.

thus less likely to benefit from a household buffer [7, 34]. In this way, the household should lead to an accumulation of (dis)advantages and to a strengthening of the society's system of stratification.

2.3 Welfare State Buffer

A further mechanism for buffering income losses is welfare state support via transfers and taxes. Germany and the US greatly differ in the generosity of the transfers and the progressivity of the taxation system.

The main state program directed at cushioning the economic consequences of job loss is unemployment insurance. A corporatist form of insurance characterizes Germany. The entitlement to benefits is based on previous labour market participation, and benefits are related to previous earnings—in this way, the welfare state also stratifies [21, 41, 44]. Also in the US eligibility to unemployment insurance is mainly related to previous employment (but other criteria apply across states) and the amount of the benefit is calculated based on previous earnings [44]. However, Germany and the US present considerable differences in the net replacement rate of unemployment benefits which has been estimated at 74% and 56% respectively [35, 40]. Overall, unemployed individuals are supported by the welfare state to a greater extent in Germany than in the US, in both the short run (within the first year) and the longer run (over a period of 5 years).

Alongside unemployment insurance, the state may intervene also through social assistance aimed at providing social protection for people in need. Social assistance is not in general directed at covering specific risks, and includes means- or income-tested benefits and minimum income protection [2]. The public system of benefits in Germany includes general social assistance and unemployment assistance. The first is a benefit directed to all residents based on a needs test [1]. The second is a means-tested benefit directed to those that are no longer entitled to unemployment insurance, and was abolished in 2005.[4] In the US, social assistance programs are basically represented by food stamps and public assistance restricted to households with children [18, 40].[5] In addition, another measure intended to support the income of low-wage workers is the Earned Income Tax Credit.

Over time, both countries have experienced changes in their policies. However, in comparative terms, the US has been and continues to be the least generous welfare state in both social assistance and unemployment benefits—and in both amount and duration of benefits. Therefore, I expect a larger welfare state buffer for Germany than for the US (*Hypothesis 4*).

---

[4] From 2005, the Hartz reform introduced 'Unemployment Benefits II', an unemployment benefit directed to those who exhausted their Unemployment Insurance entitlement or that were not eligible, and that replaced the former unemployment assistance as well as part of general social assistance.

[5] These programs include the Aid to Families with Dependent Children program and the Temporary Assistance for Needy Families program that replaced the first in 1996.

Moreover, while unemployment insurance should further stratify the income losses associated with unemployment, the targeted character of social assistance programs operates in the opposite direction suggesting that the welfare state buffer is likely to be more effective for the less-educated and thus least-resourceful individuals. Therefore, I expect that the higher the level of education, the lower the role of the welfare state (*Hypothesis 5*).[6]

3 Data and Methods

The data I use is the Cross National Equivalent File (CNEF) [24]. The CNEF includes data for Germany and the US that come from the Socio-Economic Panel (SOEP) and from the Panel Study of Income Dynamics (PSID) respectively. This file is particularly suitable because it provides longitudinal and harmonized information at the individual and household level.

The PSID started in 1968 on an annual basis but, since 1997, individuals have been interviewed biannually.[7] The SOEP started in 1984 and provides annual information for the entire time span covered. In order to have a common observational window I select the waves starting from 1984 up to 2015 for both countries. I select all individuals from 25 to 54 years old designated as the household head or his/her partner. This age selection has been chosen to include in the sample individuals that are potentially in the labour market and that may form a household. I then delete all individuals for which information about employment and incomes are missing and select only those who were in the panel for at least 5 years. Finally, I exclude from the analysis those who never entered in employment.

In line with the discussion above, I perform separate analyses for men and women as well as by levels of education.

3.1 Definition of Measures and Variables

Job loss consequences and the buffering capacity of the household and the state are captured by comparing different income concepts [14]. The first concept is *individual labor earning* (*labor income* for simplicity). This income concept includes wages and salary from employment including training, primary and secondary jobs, and self-employment, plus income from bonuses, over-time and profit-sharing [27]. This concept measures only the role played by the market in contributing to

[6] This should be true also because the taxation system, as well as other types of transfers – to which the least well off are more likely of having access to – may operate in a way that benefits the lowest strata the most.

[7] For this reason, results will be presented in two-year intervals: at the time of unemployment and two and four years later.

defining individual's economic resources. The second income concept, *equivalent pre-government household income* (*household income*), includes the contribution of the household via income pooling—thus adding to the previous concept the labour income of other household members and household's assets income, private transfers and private retirement income –, and economies of scale—which are accounted for by equivalising income via the OECD-modified equivalent scale. Finally, the third income concept, *equivalent post-government household income* (*disposable income*), includes also the contribution of the state adding household's public transfers and social security pension, and subtracting total household taxes.

All the three income concepts have been deflated using the Consumer Price Index at 2010 value, to make them comparable over time. Before adjusting these income measures, I top-coded them at the 99th percentile to avoid extreme income values.

I define the job loss event as the transition from employment to unemployment from one year to the next. Then I construct income trajectories around the event. For employment status, I decided to retrieve monthly information from the original SOEP and PSID data and merge this information to the harmonized dataset. Variables available in the original datasets provide more precise and reliable information. They indeed report individual's monthly employment status permitting in this way to precisely identify unemployment episodes and to distinguish unemployment from other types of non-employment. Based on this information, individuals are considered unemployed if they are unemployment for at least 3 months during the year.

Income trajectories are presented in relative terms. Measuring income changes in percentage terms enables exploration of whether the severity of the losses depends on prior standards of living and betters allows evaluation of unemployment consequences for individuals and households that are in different positions over the stratification ladder.

Based on these relative measures, I can disentangle the buffering capacity of the household and the state by comparing income changes across the three abovementioned income concepts. Following DiPrete and McManus [14] and Ehlert [20], I measure the household buffer as the difference between the loss in labour income and the loss in family income. In a similar way, the welfare state effect is measured as the difference in income changes between family and disposable income.[8]

3.2 The Model—The 'Distributed' Fixed-Effect

In order to model income trajectories after job loss, I use a particular specification of fixed-effects model: The distributed fixed-effects model [16, 36, 45].

[8] I decided to measure the household and welfare state buffers in absolute terms rather than in relative terms. While using a relative measure for the buffers would tell us what is the share of the total buffered loss that comes from each institution, it would not allow evaluating variations in the observed buffering capacities across social strata – which is instead possible with an absolute measure.

A general income equation can be written as

$$y_{it} = \sum_{l} \alpha_l Z_{li} + \sum_{j} \beta_j X_{jit} + \gamma EMP_{it} + u_i + \varepsilon_{it} \tag{1}$$

where y_{it} is a measure of income for individual i at time t, Z_{li} is a vector including the intercept and time-constant covariates, X_{jit} is the vector for time-varying covariates other than the employment status, EMP_{it} stands for the employment status while u_i and ε_{it} are respectively an individual-specific error term representing the effect of unobserved correlates and an idiosyncratic error term. i, l, j and t index over individuals, the observed time-constant and time-varying covariates, and time periods, respectively.

Averaging the covariates for each individual over time,

$$\overline{y}_i = \sum_{l} \alpha_l Z_{li} + \sum_{j} \beta_j \overline{X}_{ji} + \gamma \overline{EMP}_i + u_i + \overline{\varepsilon_i} \tag{2}$$

and subtracting this from Eq. (1), one obtains

$$y_{it} - \overline{y}_i = \sum_{l} \alpha_l (Z_{li} - Z_{li}) + \sum_{j} \beta_j (X_{jit} - \overline{X}_{ji}) + \gamma (EMP_{it} - \overline{EMP}_i) + (u_i - u_i) + (\varepsilon_{it} - \overline{\varepsilon_i}) \tag{3}$$

From here, the unobserved fixed-effect $\overline{u}_i$ as well as the vector containing the intercept and the time-constant covariates Z_{li} can be removed.

Now, starting from the equation of the fixed-effects model it is possible to formalize the distributed fixed-effects model by disaggregating the employment status over time, i.e. by substituting the dummy variable EMP_{it} in Eq. (3) with a set of dummy variables EMP_{pit} as in Eq. (4), where p is the number of years before unemployment if negative, and the number of years after unemployment if positive, while s represents the maximum horizon in years backward and forward from time to unemployment.

$$\sum_{p=-s}^{s} EMP_{pit} = EMP_{it} \tag{4}$$

Such model specification allows to estimate 'distributed effects': Job loss effects for each year separately, on a process time axis defined by the labor market event. The time axis is thus centred on the event. To be more precise, individuals are coded with 0 when they experience the transition from employment to unemployment; they are coded between − 1 and − 5 in the years before the transition to unemployment; and are coded between + 1 and + 5 in the years after, irrespective of their status. Observations which are 5 or more years before the event represent the reference category. In line with this model specification, the reference category also includes those who never experience unemployment events. Observations which are 5 or more

years after the unemployment event are grouped together. Results are presented in two-year intervals, i.e. when the individuals experience unemployment (time 0), and 2 and 4 years after.

Multiple episodes are dealt as follow: I consider an unemployment transition to be a 'new' unemployment transition if the individual is observed as employed for at least 4 years between the previous and the current unemployment spell (i.e. 5 years after the individual was first observed unemployed). In addition to the distributed employment status, the models control for whether the individual has a partner, number of children younger than 14, age, age squared, and year dummies.

4 Results

I turn now the attention toward the core of this paper. Here, I present income trajectories separately for levels of education and at different points in time. I start by discussing results for German men and women and I then move to American men and women. I will discuss the results in comparative perspective in the conclusions.

4.1 Germany

Figure 1 presents income trajectories for individual labour income, stratified by educational level (income trajectories for household income and disposable income are reported in the Appendix, Figs. 5, 6). Income trajectories are expressed in terms of percentage income losses with respect to the reference year before job loss.

The first row in Fig. 1 reports income trajectories in men's labour income and it shows very different income trajectories between groups, where the highest educated experience the lowest income penalty following unemployment. In the year of job loss, German men with less than high school experience a reduction of 60 percent with respect to income before job loss. Those with high school education lose 50 percent of income while those with more than high school lose 44 percent.

This signals an accumulation of disadvantages: Less educated men have the lowest levels of income, experience the highest risk of unemployment, and suffer from the largest income losses with the event (see Tab. A1 in the Appendix). Moreover, they also have more difficulties recovering from their income losses in the years after the event. Two years after job loss, income losses among the least educated are almost twice the income losses experienced by the most educated (52% vs 27%). Overall, the higher the level of education, the smaller the impact of unemployment. This is also true looking at four years after job loss (43% vs 17% comparing the least and the most educated).

These results confirm my first hypothesis: Because of the German labour market structure (and the educational system) we observe a larger penalty in the years after the event for less educated with respect to high educated individuals.

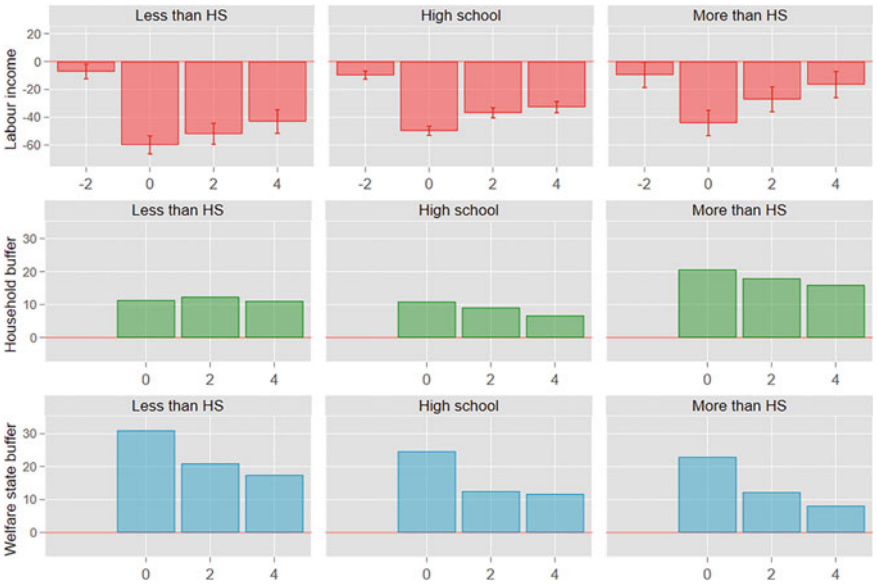


Fig. 1 Estimated income trajectories for individual labor income, and Household and Welfare State buffers, by education. Germany, men

Results for German women (Fig. 2) largely resemble those for German men: The higher the level of education, the lower the income losses. The only exception in the pattern is that the highest educated experience a similar loss at the time of the event (53%) as compared to high school educated women (51%). The least educated experience instead an income loss of 61 percent. If we look at the years following the event, we observe monotonically decreasing losses while education increases. For example, four years after the event, income losses are of 39, 29 and 24 percent respectively for low-, middle- and high-educated women. Accumulation of disadvantages seems to be at work also for women, confirming *Hypothesis 1* also for them.

Concerning the role that the household plays in redistributing resources and in managing social risks, the second row of Fig. 1 reports the 'household buffer' measured as the difference in income losses between labour income and family income. For German men, the household redistributive capacity reduces income losses. However, the household only plays a moderate role. Overall, when the event is experienced, the reduction in income loss ranges between 11 percentage points for individuals with high school or less education and 21 percentage points for those with more than high school education.

A different picture emerges if we focus on women (Fig. 2). For them, in fact, the household plays a larger role and, most importantly, its role decreases with education. At the time of the event, for example, the household buffer decreases from 31 percentage points for the least educated to 22 percentage points for the most educated.

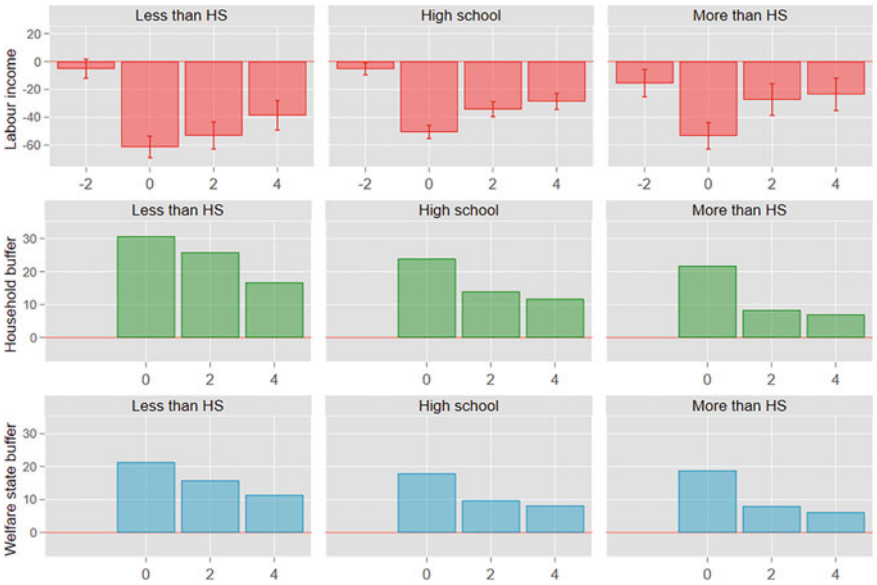


Fig. 2 Estimated income trajectories for individual labor income, and Household and Welfare State buffers, by education. Germany, women

Gender differences in the magnitude of the household buffer support *Hypothesis 2a*. On the one hand, men have lower chances to benefit from a partner's income because of the lower labour market participation of women and, on the other hand, when she is employed, she is often employed with a part-time job and thus with a lower capacity to compensate his income loss.

Concerning differences between levels of education, the household buffer is more effective for the most educated/advantaged men, in line with *Hypothesis 3*. The uneven role of the household persists also in the years that follow job loss. These patterns suggest that the household contributes to strengthening the system of socioeconomic stratification. But contrary to expectations we observe the opposite patterns for women for which the household buffer become smaller as education increases. A possible factor that could at least partly explain these observed patterns is the largest share of household income that high-educated women earn. In fact, while the share of household income attributable to man's earnings is very similar across levels of education, the share of household income attributable to woman's earnings varies substantially, increasing with education. This implies that when a high-educated woman experience unemployment, the household loses a larger share of household income with the consequence that the household will have a reduced capacity to buffer the loss. However, while this explanation can contribute to account for the observed patterns, it can hardly completely explain them; further investigation will be instead needed.

Finally, the state buffer is shown in the third row of Figs. 1 and 2. For men, we note a considerably improved situation once the state enters into play. Unemployment insurance and social transfers more generally, as well as the system of taxation make a big difference for their economic situation. For the lowest educated individuals, the German welfare state reduces income losses by up to 31 percentage points at the time of the event and, albeit with a decreasing capacity, it continues to work also four years later. The state's role, however, decreases the more we move toward the highest levels of education: At the time of the event it is 25 and 23 percentage points for middle- and high-educated, respectively.

Patterns for women are similar, although differences between levels of education are less marked, especially at the time of the event: The welfare buffer is 21, 18 and 19 percentage points for low-, middle- and high-educated women, respectively. Results overall corroborate *Hypothesis 5* for both men and women: The state plays an important role in first mitigating and then equalizing the negative consequences of job loss for individuals' income.

Existent stratification is strengthened by the market and then alleviated by the state. The household plays a mixed role contributing to strengthening inequality for men but reducing inequality for women. As a result of these complex processes, we observe very similar trajectories in disposable income across levels of education for women. On the contrary, with observe different trajectories for mem, with the most educated not experiencing any penalty in disposable income.

4.2 United States

I now turn my attention towards the United States. At the time of the event, men lose between 44 (low-educated) and 47 (mid- and high-educated) percent of their labor income (Fig. 3). In the following years, the losses decrease only slightly and four years later they range between 30 (mid-educated) and 36 percent (high-educated).

Concerning differences across educational levels, results support *Hypothesis 1*. In the US ILMs, income losses after job loss do not vary significantly across levels of education.

Income trajectories for American women, presented in Fig. 4, are rather similar to those for men. However, we observe a larger penalty for the less educated groups in the year of job loss (55 percent), although differences are contained—by 3 and 7 percentage points compared to mid- and high educated respectively.

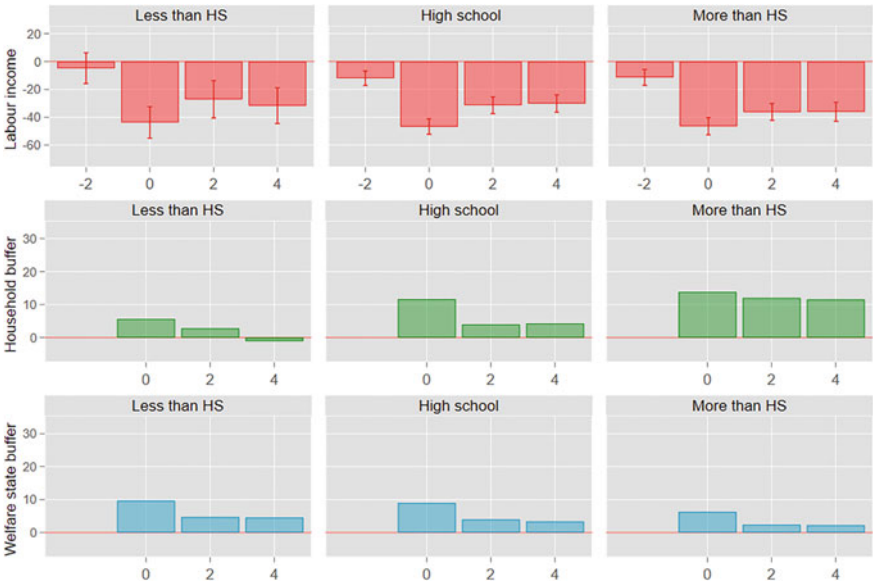


Fig. 3 Estimated income trajectories for Individual Labor Income, and Household and Welfare State buffers, by education. United States, men

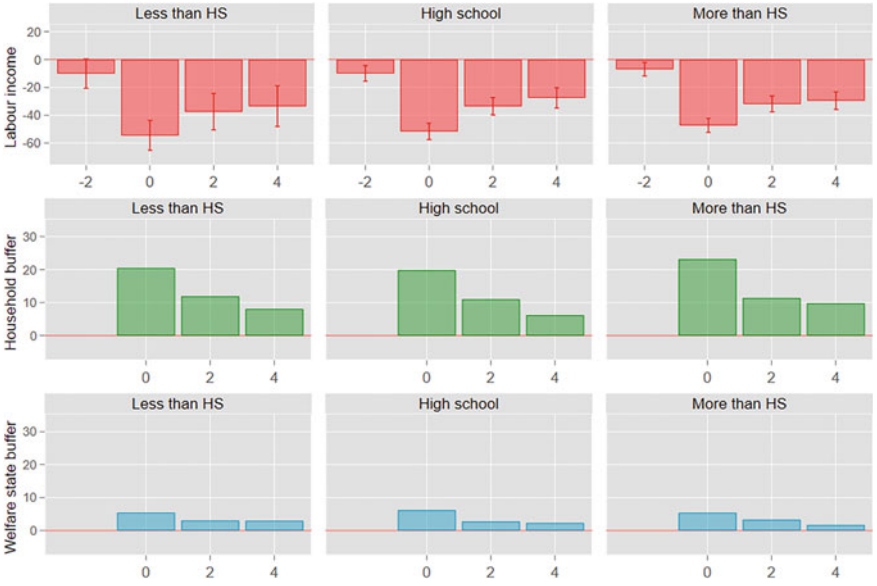


Fig. 4 Estimated income trajectories for Individual Labor Income, and Household and Welfare State buffers, by education. United States, women

Once we consider the role of the household, the picture changes for men and even more so for women (see Figs. 7, 8 for income trajectories). The household reduces men's losses to a limited extent between 6 and 14 percentage points, with the most educated benefitting the most, both at the time of the event and afterwards. *Hypothesis 3* is thus supported: The household tends to reinforce pre-existing (dis)advantages.

Hypothesis 2a is also corroborated. The household role is considerably larger for women compared to men. For example, among low-educated at the time of the event, the household buffer for women is more than three times the buffer for men (21 vs 6), while for high-educated it is about 10 percentage points higher (23 vs 14). However, I find only limited support to *Hypothesis 3* for American women: The household buffer tends to be very similar across groups, with the exception that we observe a slightly larger buffer for the most educated individuals at the time of the event compared to the other groups. As a result, educational inequalities in income trajectories have only slightly changed with respect to individual labor income.

Finally, when also the state is at work, income trajectories change again, although only partially. As expected, the residual welfare state that characterizes the United States does not contribute much to manage social risks. Concerning American men, the state mitigates income losses by no more than 10 percentage points at the time of the event. Over time, the state buffer is still present, although to a negligible extent. The same patterns hold for women, although they benefit from slightly less state support. *Hypothesis 5* cannot be fully corroborated. Only for men the role of the state decreases with education, where we observe a buffer of 10 percentage points for low-educated while a buffer of 6 percentage points for high-educated men. Overall, and differently from Germany, in the United States unemployment insurance and social assistance play a much-reduced role in supporting unemployed individuals.

5 Conclusions

This paper set out to investigate the economic consequences of job loss in international comparison and among individuals with different educational levels. I contribute to the existing literature by systematically considering events, attributes (in terms of level of education and sex), and their interaction in context, thus adopting comparative, life-course, and stratification perspectives. Economic consequences are conceptualized as the result of the interaction between the event, the individual position within the socioeconomic hierarchy, and the broader context. I find some evidence of the accumulation of inequalities over the life-course, combined with different capacities of welfare institutions to mitigate the negative economic consequences of job loss.

Education affects the negative consequences of job loss and may exacerbate the disadvantage of the already less advantaged groups. Differences between levels of education for market income are clearly visible for German men and women—for the latter especially in the trajectory after the event. Stratification of income trajectories by education is much less visible in the US. These patterns are in line with the different character that education has in OLMs and ILMs, as proposed in *Hypothesis 1*.

A smaller income loss is observed in the US than in Germany, which might be attributed to the higher dynamism of the US labour market where individuals may experience shorter unemployment episodes. Obviously, the main channel through which unemployed people can recover their previous income levels is re-employment.

The results show a substantial contribution of the household in managing the consequences of unemployment. In both Germany and the US, women benefit the most from the support of other household members, mainly the partner, this supporting *Hypothesis 2a*. Comparing men in the two countries, American men are not supported by their partner to a greater extent than German men (*Hypothesis 2b* is not supported), notwithstanding the higher labour market participation and labour market intensity of women in the US. In fact, results show a slightly larger buffer for German men, which might be explained by the higher share of singles in the US (Tables 1, 2). Both men and women in the US are more likely than Germans not to have a partner who supports them when facing adverse life-course events. This might also explain the lower household buffer observed for the US women compared with German women.

Regarding the role of the state, the welfare state reduces income losses in both countries. As expected (*Hypothesis 5*), the welfare state buffer tends to be inversely related to individuals' education—the largest buffer for the lowest educated/least resourceful individual could most likely be attributed to the targeted character of social assistance programs and the progressivity of the taxation system. In addition, the role of the state is much more prominent in Germany than in the United States (confirming *Hypothesis 4*). The more generous German welfare state performs much better in managing the negative consequences of job loss.

Comparing the findings with previous research, overall my analyses are in line with DiPrete and McManus [14] and Ehlert [20] regarding the (un)equalizing role of the household and the state. However, concerning the stratification of income trajectories, I find clearer patterns of stratification between groups in Germany as compared to [20]. I attribute this difference to the fact that education, as compared to household income, is better suited to capture inequality between strata in the German labor market.

Overall, my results show that institutions play a substantial role in shaping income trajectories of individuals. However, the extent to which institutions contribute to fostering or mitigating the income losses associated with unemployment varies according to several aspects, including gender. While women are those who benefit the most from the household, men are better sheltered by the welfare state.

Institutions in both countries play a considerable role in strengthening or containing inequality and in (re)producing the system of socioeconomic stratification. These patterns are especially clear-cut for men.

The market emerged to operate as an inequality 'booster' in Germany but much less so in the US. The role of the household points toward the same direction by supporting the highest educated the most, with the exception of German women. By contrast, the state operates in the opposite direction by targeting its intervention to the least educated.

In terms of final welfare, as captured by household disposable income, individuals in the US are those who lose the most, with a similar magnitude across levels of education. In fact, the smaller household and especially welfare state buffer in the US do not replace income losses to the same extent as in Germany.

In terms of inequality, the three institutions tend to counterbalance one another with the result that losses in disposable household income are rather even across educational groups (Figs. 5 and 8). The exception in this complex combination of forces are German men. For them, while the equalizing role of the state almost perfectly counterbalance the un-equalizing role of the household, initial educational inequality in the labour market translates into inequality in final welfare as captured by disposable household income.

Acknowledgments The author acknowledges the support of CRITEVENTS. The CRITEVENT project is financially supported by the NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course, which is co-funded by the European Commission through Horizon 2020 un-der grant agreement No. 724363

Appendix

Table 1 Individual characteristics comparing those always employed and those that have experienced at least one episode of unemployment, Germany

	Men							
	Always employed				With unemployment experience			
	Less than HS	High School	More than HS	Total	Less than HS	High School	More than HS	Total
Age	42.4	41.19	41.83	41.47	39.9	40.98	40.53	40.74
East-Germany	2.77	20.11	15.14	17.24	9.65	42.29	17.89	34.05
With partner	89.08	88.17	87.43	88.03	87.46	81.99	77.06	82.47
HH size	3.61	3.19	3.20	3.23	3.61	3.05	2.83	3.13
Mean labor inc	36,485	42,982	62,761	48,324	31,422	32,308	49,740	34,115
Unemployed					30.67	25.01	16.1	25.13
Individuals	803	5,587	2,611	9,001	336	1,323	186	1,841
Episodes					356	1,408	194	1,958
Person years	7,014	54,059	25,753	86,826	3,877	15,748	2,180	21,805
	Women							
	Always employed				With unemployment experience			
	Less than HS	High School	More than HS	Total	Less than HS	High School	More than HS	Total
Age	43.43	41.20	40.82	41.31	41.50	41.37	39.42	41.14
East-Germany	6.46	23.51	25.36	22.36	17.14	45.49	24.00	36.76
With partner	83.97	79.30	77.81	79.32	76.30	77.16	68.07	75.78
HH size	3.25	2.90	2.92	2.94	3.24	2.90	2.60	2.93
Mean labor inc	21,464	26,019	35,828	28,505	20,983	20,555	28,934	21,842
Unemployed					26.37	25.32	18.07	24.59
Individuals	798	4,550	2,159	7,507	402	1,245	258	1,905
Episodes					421	1,311	271	2,003
Person years	6,624	39,689	19,840	66,153	4,334	13,773	2,750	20,857

Table 2 Individual characteristics comparing those always employed and those that have experienced at least one episode of unemployment, United States

	Men							
	Always employed				With unemployment experience			
	Less than HS	High School	More than HS	Total	Less than HS	High School	More than HS	Total
Age	42.4	41.19	41.83	41.47	39.9	40.98	40.53	40.74
Race								
White	57.54	67.77	78.23	72.58	41.89	56.81	70.74	60.70
Black	33.51	28.57	17.11	22.7	49.48	39.4	23.7	34.02
Other	8.95	3.66	4.66	4.72	8.63	3.79	5.55	5.27
With partner	85.1	85.5	85.54	85.49	70.27	74.74	74.89	74.15
HH size	3.58	3.33	3.27	3.32	3.26	3.18	3.02	3.12
Mean labor inc	34,527	46,106	69,272	57,818	26,104	34,746	54,762	42,534
Unemployed					22.06	16.86	14.4	16.55
Individuals	501	1,673	2,503	4,677	288	691	678	1,657
Episodes					304	757	748	1,809
Person years	4,369	16,093	25,339	45,801	2,607	7,360	7,742	17,709
	Women							
	Always employed				With unemployment experience			
	Less than HS	High School	More than HS	Total	Less than HS	High School	More than HS	Total
Age	41.55	39.65	38.42	39.1	39.01	38.22	38.26	38.33
Race								
White	50.6	64.08	68.54	65.57	33.23	46.48	55.88	49.12
Black	43.36	31.36	26.87	29.74	62.39	49.36	40.03	46.73
Other	6.04	4.56	4.59	4.69	4.38	4.16	4.09	4.16
With partner	62.84	72.72	72.29	71.74	44.66	58.93	62.62	58.86
HH size	3.53	3.25	3.08	3.18	3.57	3.27	3.03	3.20
Mean labor inc	19,578	27,247	40,482	34,085	16,827	22,410	31,052	25,764
Unemployed					23.29	16.04	15.07	16.48
Individuals	397	1,644	2,505	4,546	279	771	815	1,865
Episodes					296	849	896	2,041
Person years	3,095	15,164	23,227	41,486	2,284	7,972	8,483	18,739

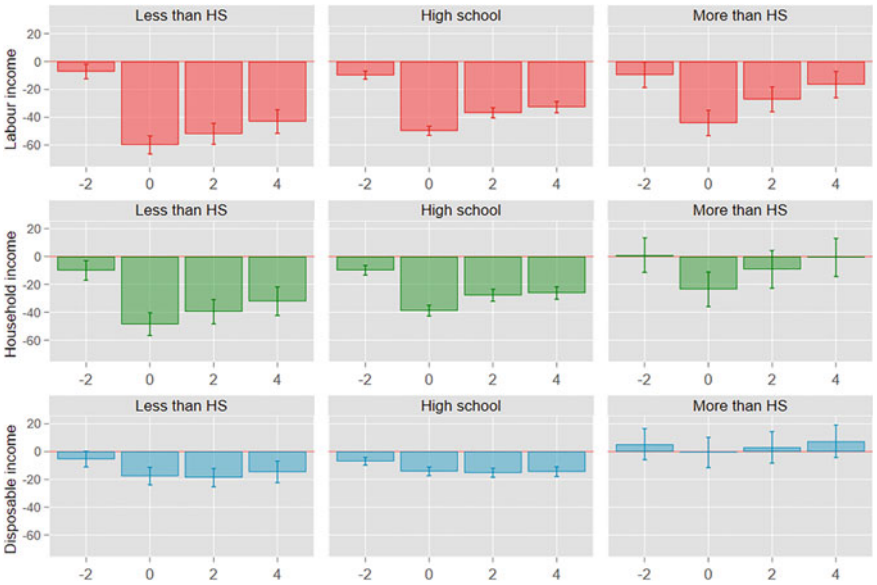


Fig. 5 Estimated income trajectories for different income concepts at different points in time, by level of education. Germany, men

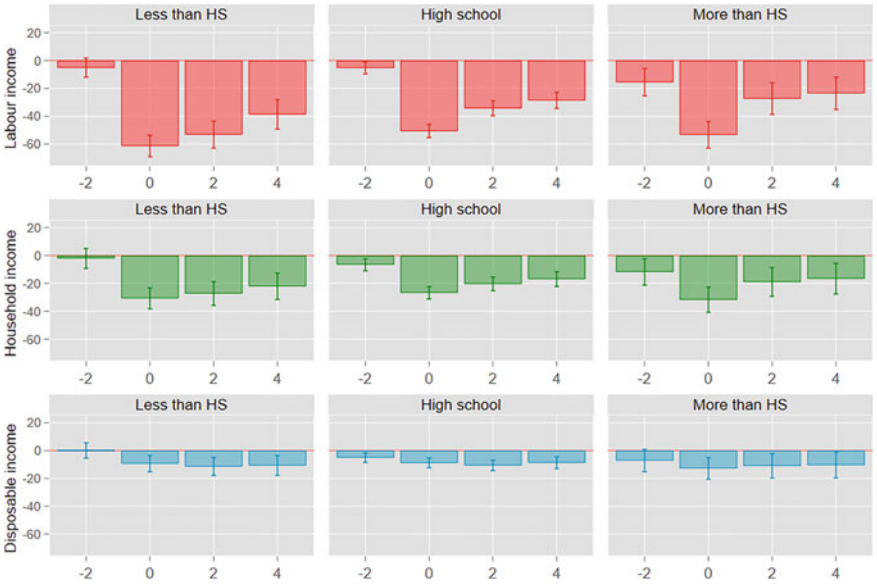


Fig. 6 Estimated income trajectories for different income concepts at different points in time, by level of education. Germany, women

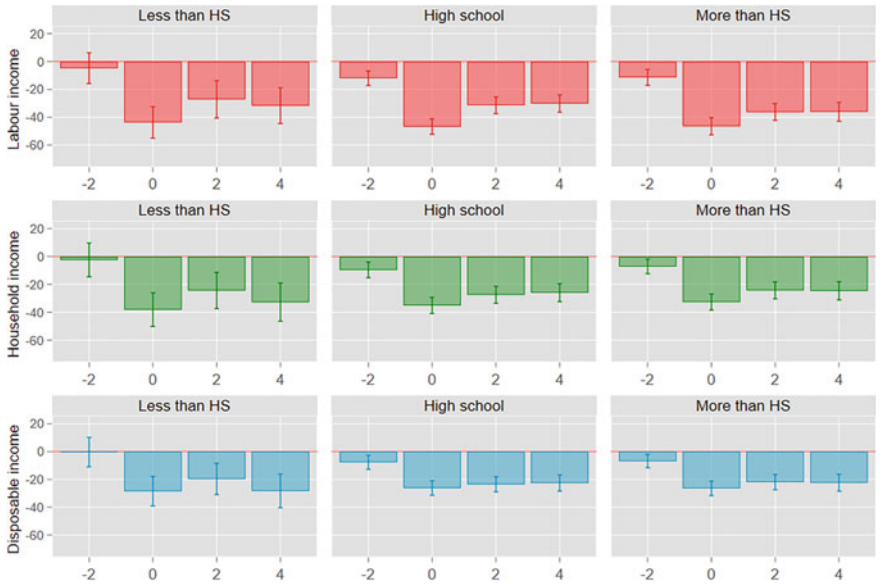


Fig. 7 Estimated income trajectories for different income concepts at different points in time, by level of education. United States, men

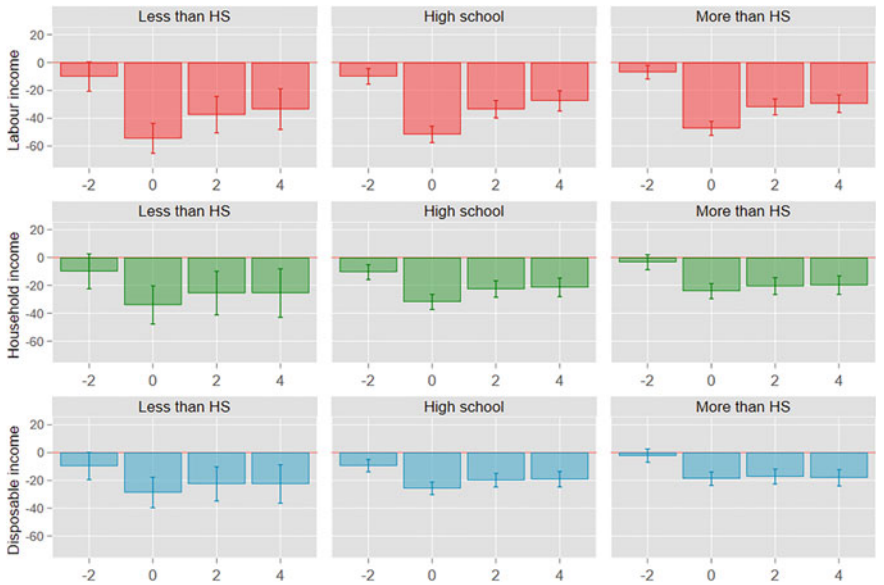


Fig. 8 Estimated income trajectories for different income concepts at different points in time, by level of education. United States, women

References

1. Adema, W., Gray, D., Kahl, S.: Social assistance in germany. OECD Publishing (2003). https://doi.org/10.1787/338133058573
2. Bahle, T., Pfeifer, M., Wendt, C.: Social assistance'. In *The Oxford Handbook of the Welfare State*, edited by Castles, F. G., Leibried, S., Jane, L., Obinger, H., Pierson, C. 448–61 (2010)
3. Blossfeld, H.P., Timm, A.: *Who Marries Whom? Educational Systems as Marriage Markets in Modern Societies*. Edited by Hans Peter Blossfeld, Timm. A., Dordrecht: Kluwert (2003)
4. Blossfeld, H.P., Drobnic, S.: Careers of couples in contemporary societies: from male breadwinner to dual-earner families. Oxford University Press (2001)
5. Brand, J.E.: Enduring Effects of Job Displacement on Career Outcomes. Univ. Wis.-Madison, Madison, WI (2004)
6. Breen, R.: Risk, recommodification and stratification. Sociology **31**(3), 473–489 (1997). https://doi.org/10.1177/0038038597031003006
7. Brülle, J.: Demographic trends and the changing ability of households to buffer poverty risks in Germany. *European Sociological Review* **32**(6): 766–78 (Jan. 2016) https://doi.org/10.1093/esr/jcw033.
8. Burda, M.C., A. Mertens.: Estimating wage losses of displaced workers in Germany. *Labour Economics* **8**(15–41) (2001)
9. Burkhauser, R.V., Duncan, G.J., Hauser, R., Berntsen, R.: Economic burdens of marital disruptions: a comparison of the united states and the federal republic of Germany. Rev. Income Wealth **36**(4), 319–333 (1990). https://doi.org/10.1111/j.1475-4991.1990.tb00317.x
10. Crompton, R.: Employment and the Family. The Reconfiguration of Work and Family Life in Contemporary Societies. Camb. Univ. Press., Cambridge (2006)
11. Cutuli, G., Grotti, R.: Heterogeneity in unemployment dynamics: (Un)observed drivers of the longitudinal accumulation of risks. Res. Soc. Strat. Mobil. **67**(June), 100494 (2020). https://doi.org/10.1016/j.rssm.2020.100494
12. DiPrete, T.A.: Life course risks, mobility regimes, and mobility consequences : a comparison of Sweden, Germany, and the United States 1. Am. J. Sociol. **108**(2), 267–309 (2002)
13. DiPrete, T.A., Eirich, G.M.: Cumulative Advantage as a mechanism for inequality: A review of theoretical and empirical developments. Ann. Rev. Sociol. **32**(1), 271–297 (2006). https://doi.org/10.1146/annurev.soc.32.061604.123127
14. DiPrete, T.A., McManus, P.A.: Family change, employment transitions, and the welfare state: household income dynamics in the United States and Germany. Am. Sociol. Rev. **65**(3), 343 (2000). https://doi.org/10.2307/2657461
15. Doeringer, P. B., Piore, M. J.: Internal labor markets and manpower analysis. Edited by heath lexington books. lexington (1971)
16. Dougherty, Christopher.: The marriage earnings premium as a distributed fixed effect. *Source: The Journal of Human Resources* **41**(2): 433–43 (2006).
17. Drobnic, S., Blossfeld, H.-P., Rohwer, G.: Dynamics of women's employment patterns over the family life course: a comparison of the United States and Germany. J. Marriage Fam. **61**(1), 133 (1999). https://doi.org/10.2307/353889
18. Eardley, T., Bradshaw, J., Ditch, J., Gough, I.: Social assistance in OECD countries: Synthesis Report. London (1996).
19. Ehlert, M.: Buffering income loss due to unemployment: family and welfare state influences on income after job loss in the United States and Western Germany. Soc. Sci. Res. **41**(4), 843–860 (2012). https://doi.org/10.1016/j.ssresearch.2012.02.002
20. Ehlert, M.: Job loss among rich and poor in the United States and Germany: Who Loses More Income? Res. Soc. Strat. Mobil. **32**(1), 85–103 (2013). https://doi.org/10.1016/j.rssm.2012.11.001
21. Esping-Andersen, G.: The Three Worlds of Welfare Capitalism. Polity Pre, Oxford (1990)
22. Esping-Andersen, G.: Social foundations of postindustrial economies. Oxf. Univ. Press., New York (1999)

23. Eyraud, F., Marsden, D., Silvestre, J.J.: Occupational and Internal labour markets in Britain and France. Int. Labour Rev. **129**(4), 501–517 (1990)
24. Frick, J.R., Jenkins, S.P., Lillard, D.R., Lipps, O., Wooden, M.: The Cross-national equivalent file (CNEF) and its member country household panel studies. Schmollers Jahr. **127**, 627–654 (2007)
25. Gangl, M.: Welfare states and the scar effects of unemployment: a comparative analysis of the United States and West Germany1. Am. J. Sociol. **109**(6), 1319–1364 (2004). https://doi.org/10.1086/381902
26. Gangl, M.: Scar effects of unemployment: an assessment of institutional complementarities. Am. Sociol. Rev. **71**(6), 986–1013 (2006). https://doi.org/10.1177/000312240607100606
27. Grabka, M.: Codebook for the $PEQUIV file CNEF variables with extended income information for the SOEP. Vol. 65 (2012)
28. Grotti, R., Scherer, S.: Accumulation of employment instability among partners—evidence from six eu countries. Eur. Sociol. Rev. **30**(5), 627–639 (2014). https://doi.org/10.1093/esr/jcu063
29. Grotti, R., Scherer, S.: Does gender equality increase economic inequality? evidence from five countries. Res. Soc. Strat. Mobil. **45**(September), 13–26 (2016). https://doi.org/10.1016/j.rssm.2016.06.001
30. Hall, P.A., Soskice, D.W.: Varieties of capitalism: the institutional foundations of comparative advantage. Edited by Hall P.A., Soskice, D.W. Oxford: Oxford University Press (2001)
31. Kalleberg, A.L.: Comparative perspective on work structures and inequality. Ann. Rev. Sociol. **14**, 203–225 (1988)
32. Kalleberg, A.L., Sorensen, A.B.: The Sociology of labour markets. Ann. Rev. Sociol. **5**, 351–379 (1979)
33. Kalmijn, M.: Intermarriage and homogamy: Causes, patterns, trends. Ann. Rev. Sociol. **24**(1), 395–421 (1998). https://doi.org/10.1146/annurev.soc.24.1.395
34. Kalmijn, M.: The educational gradient in marriage: a comparison of 25 european countries. Demography **50**(4), 1499–1520 (2013). https://doi.org/10.1007/s13524-013-0229-x
35. Korpi, W., Palme J.: The social citizenship indicator program (SCIP). Stockholm: Swedish institute for social research (2007)
36. Kratz, F., Brüderl, J.: 'Returns to Regional Migration: Causal Effect or Selection on Wage Growth?' *SOEPpapers*. DIW, Berlin (2012)
37. Marsden, D.: Institutions and labour market mobility: occupational and internal labour markets in Britain, France, Italy and West Germany. In *Labour relations and economic performance*, edited by Brunetta, R., Dell'Aringa, C. 414–38. Basingstoke: Macmillian. (1990)
38. McManus, P.A., DiPrete, T.A.: Market, family, and state Ssources of income instability in Germany and the United States. Soc. Sci. Res. **29**(3), 405–440 (2000). https://doi.org/10.1006/ssre.2000.0675
39. Merton, R.K.: The matthew effect in science. The reward and communication systems of science are considered. Science **159**, 56–63 (1968)
40. OECD.:Benefits and Wages 2007. Paris (2007)
41. Oesch, D.: Stratifying welfare states: class differences in pension coverage in stratifying welfare states: Class differences in pension coverage in Britain, Germany, Sweden and Switzerland. Switz. Swiss J. Sociol. **34**(3), 533–554 (2008)
42. Ruhm, C.J.: Are workers permanently scarred by job displacement. Am. Econ. Rev. **81**, 319–324 (1991)
43. Schwartz, C.R., Mare, R.D.: Trends in educational assortative marriage from 1940 to 2003. Demography **42**(4), 621–646 (2005). https://doi.org/10.1353/dem.2005.0036

44. Sjöberg, O., Palme, J., Carroll, E.: Unemployment Insurance. In *The Oxford Handbook of the Welfare State*, edited by Castles, F.G., Leibried, S., Jane, L., Obinger, H., Pierson, C. 420–34. Oxford: Oxford University Press. (2010) https://doi.org/10.1093/oxfordhb/9780199579396.003.0029
45. Yankow, J.J.: Migration, job change, and wage growth: a new perspective on the pecuniary return to geographic mobility. J. Reg. Sci. **43**(3), 483–516 (2003). https://doi.org/10.1111/1467-9787.00308

How Much Do Knowledge About and Attitude Toward Mobile Phone Use Affect Behavior While Driving? An Empirical Study Using a Structural Equation Model

Carlo Cavicchia, Pasquale Sarnacchiaro, and Paolo Montuori

Abstract Road accidents are the eighth cause of death worldwide and the first among young people, and, in turn, distraction is considered one of the main determinants of road accidents. One of the most important factors of distraction while driving appears to be the use of a mobile phone. This paper aims at simultaneously analyzing individual Knowledge, Attitudes, and Behaviors toward the use of mobile phones while driving in one of the largest and most populous metropolitan areas of Italy, Naples. The study analyzes a sample of 774 questionnaires which consist of 39 questions each. Several socio-demographic characteristics (i.e., gender, age, profession, and others) are considered within this study. A Structural Equation Model displays that the relationship between Knowledge and Behavior is not direct and it passes through the Attitude. The results of this study might be used for the creation of targeted educational programs, community-based interventions, and legal regulations.

Keywords Mobile phones · Drive · Knowledge · Attitudes · Behaviors · Cross-sectional survey · Structural equation model · Measurement model

1 Introduction

The World Health Organization (WHO) declared that traffic accidents were the eighth cause of death worldwide and the first amongst subjects aged 5–29 years, estimating 1.35 million people died as a result of road crashes in 2016. The data showed that,

C. Cavicchia (✉)
Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands
e-mail: cavicchia@ese.eur.nl

P. Sarnacchiaro
Department of Economics Management and Institution, University of Naples Federico II, Naples, Italy
e-mail: sarnacch@unina.it

P. Montuori
Department of Public Health, University of Naples Federico II, Naples, Italy
e-mail: pmontuor@unina.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_17

with an average rate of 27.5 deaths per 100,000 population, the risk was more than 3 times higher in low-income countries than in high-income countries where the average rate was 8.3 deaths per 100,000 population [7]. In the European Union (EU), 25,624 people died in road accidents in 2016; across the EU Member States, the highest number of road traffic victims were recorded in France (3,477), Italy (3,283), Germany (3,206) and Poland (3,026) [1]. Italy registered 3,325 road fatalities in 2018, corresponding to a mortality rate of 5.5 deaths per 100,000 population and representing a 1.6% decrease on the 3,378 fatalities occurred in 2017. Since 2010, road fatalities decreased across all road user groups and age categories, except for the elderly. According to preliminary data of the first semester of the 2019, from January to June, the number of road accidents with personal injury was 82,048 and the number of victims 1,505; the mortality index is 1.8 [4]. As for fatalities by road user groups, in Italy, passenger car occupants were the group most affected by road crashes. In 2018, passenger car occupants accounted for a plurality of road deaths with 43% of the total. They were followed by motorcyclists (21%), pedestrians (18%) and cyclists (7%). It is worth noticing that cars and motorcycles represented, respectively, 72% and 13% of the vehicle fleet [3], contrary to other parts of the world where riders of motorized two- or three-wheel vehicles were the most numerous [7]. According to data processed by the Italian Statistical Institute (ISTAT), the economic impact on social costs was 17 billion-euro in 2018, equivalent to 1% of Gross Domestic Product (GDP) [4].

One of the most important causes of distraction while driving appears to be the use of a mobile phone [49]. This study thus analyzes the behaviors enacted by Italian drivers regarding mobile phone use while driving, as well as their level of mobile phone involvement and its frequency of use. The key aim of this study is to analyze knowledge, attitudes, and behavior towards the use of mobile phones while driving in one of the largest and populous metropolitan areas of Italy, Naples. Around the world, several studies have been conducted on the use of mobile phones while driving, but to the best of our knowledge none of them ever analyzed knowledge, attitudes, and behaviors simultaneously. Analysis of knowledge, attitudes and behaviors about the risks of mobile phone usage while driving can lead us to identify its determinants in order to obtain the means to sensitize public opinion and improve people's awareness regarding the correct behavior to adopt while driving.

This paper is structured in several sections: a review of literature, followed by the description of materials and methods used in the study and analysis approach. Next, is reported a summary of the collected results and at the end discussion of the findings and conclusion are presented.

2 Literature Review

The use of a mobile phone while driving represents one of the main causes of distraction, even though laws take into account penalties and in some cases even driving license withdrawal in case of usage while driving (i.e., since 2002 in Italy the use of

handheld mobile phones while driving is not permitted); in 2015–16, according to the "ULISSES" monitoring system, 5.1% of the drivers used a cellphone while driving [3, 49]. Other studies showed how mobile phone distraction is a more frequent contributing factor in severe road fatalities and it is a rapidly growing problem. Mobile phone related behavior while driving is therefore an important issue to be studied in order to recognize it as a worldwide major health problem. In this section, we present a literature review about the mobile phone behaviors, attitudes and knowledge.

2.1 Mobile Phone Behavior

Mobile phone usage while driving is a worldwide trend for many drivers and it can be useful for real-time traffic updates, navigation, or emergency calls, but at the same time other applications like social networks are unnecessary and can cause further distraction. Driving simulation studies indicate that dual tasking, such as using a mobile phone while driving, can be detrimental to driving performance [13, 14, 20]. According to an Australian study, looking for more than 2 s at a mobile phone while driving is the most common and frequent habit amongst drivers [35]. Specifically, prior researches show that the visual, manual, and cognitive distractions associated with text messaging while driving could contribute to higher crash rates, especially for younger drivers [35, 44].

The International Transport Forum for Road Safety in Australia [2] reports the constant increase of the use of mobile phones while driving: about 60% of drivers use a mobile phone to read (32%) or send (18%) text messages. Another less recent Australian study found that 27% of drivers texted while driving although it was illegal [55]. In a geographical and culturally similar country, New Zealand, more than half (57.3%) of the participants used a mobile phone at least occasionally while driving [45]. A study conducted in Hanoi (Vietnam) showed that 8% of 26,360 riders use a mobile phone while driving [50]. In Qatar, 11.48% of drivers use mobile phone behind the wheel [43]. In a survey on the use of mobile phones while driving, which was carried out in Israel by Tomer-Fishman in 2010 [48], it was found that 81% reported not sending a text message in the past seven days, 48% avoided reading an incoming message, 13% read messages immediately, and 39% waited to attend to reading while the vehicle was stopped. A cross-sectional survey showed that in Ghana 96.4% of drivers knew that the law prohibits the use of mobile phones while driving but the majority (59.6%) did not routinely comply with the law; among drivers who reported phone use while driving, 44.6% stated they used the hands-free feature [16]. A research note from the US Department of Transportation in 2018 showed that only 3.2% of North-American drivers talked on handheld phones, a percentage that increased from 2.9% in 2017 [6]. In 2015, Huisingh et al. [25] revealed that 31.4% of drivers talked on the phone and 16.6% texted or dialed. In a South-American country, Colombia, Oviedo-Trespalacios and Scott-Parker [36] found that 78% of drivers aged 15–25 years old used a mobile phone at least occasionally when driving.

In Europe, studies conducted in different countries report contrasting results on the use of mobile phones while driving with percentages ranging between 9% and 81% and averages around 30%. A recent study conducted in Ukraine, for example, found out that almost a third of the people interviewed reported using their phone on a daily basis to write (22.2%) or read (38.2%) text messages while driving [23]. In the United Kingdom, almost 30% of study participants reported answering calls while driving on a daily basis or more [46]. In a sample from a Spanish university, a research found that more than 60% of respondents used a mobile phone while driving and that the phone was mostly used for making calls, rather than using SMS [19]. A study conducted by Pöysti et al. [39] reported that 81% of Finnish drivers used their phone in the car at least sometimes, with 9% of them using it over 15 min a day and 44% phone-using drivers admitted having experienced hazardous situations in using a phone. In Italy, few studies were conducted on the use of mobile phones while driving. However, a noteworthy recent study conducted by Valent et al. [51] in Udine (Northern Italy) reported that the prevalence of mobile phone use was 9.9% among drivers waiting at red traffic lights and 6.5% among those moving along the streets; also the type of use was recorded: prevalence of texting was 7.2% at traffic lights and 5.0% in moving vehicles, prevalence of phone calls was 3.3% and 3.6%, respectively. Moreover, Gariazzo et al. [18] showed positive associations between road crashes rates and the number of calls, texts, and internet connections, with incremental risks of 17.2, 8.4 and 54.6% per increases of 5 calls/100 people, 3 text/100 people, and 40 connections/100 people, respectively, detecting small differences across cities. Another less recent study conducted in Florence from 2005 to 2009 reported that the average mobile phone use while driving was 4.5% [28]. Beyond the aforementioned aspects, it is necessary to investigate the causes of drivers' behaviors and try to understand which factors induce these distracting conducts. Different researches carried out on the subject explored the frequency of mobile phone usage while driving [53] or the psychosocial factors associated with mobile phone use while driving [23]; but the analysis of the behaviors related to knowledge and attitudes at the same time is still an open and compelling research field.

2.2 Knowledge, Attitudes and Behaviors

The literature research demonstrates that behaviors are the results of knowledge, attitudes, or their interaction. Several studies indeed—not linked to the mobile phone usage while driving—analyze practices, actions and conducts as functions of knowledge and attitudes [5]. UNICEF, for example, is conducting a knowledge, attitude, and practice survey on COVID-19 (2020). De Pretto et al. [15], in a study conducted in Malaysia on the link between atmospheric haze pollution and outdoor sports, showed that higher levels of knowledge and concerned attitudes translate into a greater likelihood of engaging in protective practices. Another recent study about smoking knowledge, attitude and practice in Dubai concluded that the majority of never and ex-smokers had good knowledge level and positive attitude toward

anti-smoking statements [10]. Furthermore, the simultaneous analysis of knowledge, attitudes and practices was also performed in other two Italian studies. The first—carried out in healthcare personnel about hand decontamination—showed that the positive attitude was significantly higher among older and female personnel and in those with a higher level of knowledge [34]. The second study demonstrated that Genetically Modified Foods (GMF) consumption in Italian students depended on the knowledge of the impact of GMF on health and the environment [30].

To the best of our knowledge, no study ever analyzed knowledge, attitudes, and behaviors simultaneously in context of the mobile phone use while driving. In fact, Nevin et al. [33] analyze simultaneously knowledge, attitudes and behaviors on the use of the mobile phone while driving, but the survey refers exclusively to police officers and a very limited cohort (i.e., only 26 participants). Although Adeola et al. [8] analyzed at the same time knowledge, attitudes and behaviors in a sample of 1,238 teenagers, the knowledge referred only to the effects of an educational program carried out before and after the survey. In 2011, Hassen et al. [22] conducted a quantitative cross-sectional study with a sample size of 350 drivers—in detail, 75 taxi, 103 Baja and 172 private owned car—but the knowledge refers only to the meaning of 10 road signs. Moreover, another key limitation of the study was that the majority were males (96.9%). Finally, the simultaneous study of knowledge, attitudes and their interactions to behaviors in order to develop health education and community-based interventions appears crucial to develop knowledge and positive changes regarding the attitudes related to mobile phone use while driving.

3 Material and Methods

In this section we present the specifications of surveyed participants, the procedure we followed and the characteristics of the questionnaire Sect. 3.1, along with the main aspects of the statistical analysis we carried out Sect. 3.2.

3.1 Participants and Procedure

A cross-sectional, survey-based study was employed. From the beginning of June 2019 until the end of January 2020, we surveyed adults in the entirety of the metropolitan city of Naples, Italy, through a questionnaire (available upon request from the corresponding author). From the beginning of June 2019 until the end of January 2020, we surveyed adults in the entirety of the metropolitan city of Naples through a questionnaire (available upon request from the corresponding author). The sampling framework for inclusion in the study was that participants had a driver's license, a smartphone, and resided in the metropolitan area of Naples. Participants were recruited from a snowballing of the researchers' families and friends. Snowball sam-

pling was used to include participants in a wider population to increase the representativeness of the sample.

The questionnaire was anonymous, and no personally identifiable information was collected. At the time of filling in, the questionnaire was explained verbally to each participant, exposing the aim of the study and that the data collected would respect privacy and anonymity. The questionnaire consists of demographic information about the participant (age, gender, type of driven vehicle, education level, profession, years of driving and smoke) and three pools of queries focusing on knowledge, attitudes and behaviors concerning the habit and frequency of mobile phone use while driving, for a total of 46 questions. Knowledge and attitudes were assessed on a three-point Likert scale with options for "agree", "neither agree nor disagree", and "disagree", while inquiries regarding behavior were presented in a four-answer format of "never", "sometimes", "often", and "always".

The data obtained were elaborated by a Factorial Analysis, using the method of minimum residual (MINRES), in order to confirm the presence of three main aspects (latent variables): Knowledge, attitude and Behaviour. This analysis has been conducted on the polychoric correlation coefficient matrix because the variables have been measured by a Likert scale with 3 and 4 ordinal categories [21]. The number of factors was chosen through the criterion based on the derivation of factors with an eigenvalue greater than one, this criterion has been compared with other methods proposed in the literature [24, 52].

Finally, all the collected questionnaires were digitalized submitting the codified answers in an Excel (MS Office) worksheet.

3.2 Statistical Analysis

The SEM is a statistical method for testing and estimating at once causal relationships among multiple independent and dependent latent (LVs) and manifest variables (MVs). SEM entails different sub-models. The structural model comprises the relationships among the LVs which have to be developed from theoretical considerations. The independent LVs are also referred to as exogenous LVs and the dependent LVs as endogenous LVs. For each of the LVs within the SEM a measurement model has to be defined. These models embody the relationships between the MVs and the LVs, and they can be either reflective or formative. In SEM related literature, two different types of techniques are established: covariance-based ones, as represented by LInear Structural RELations (LISREL, [26]), and variance-based ones, of which the Partial Least Squares (PLS) path modelling [58] is the most prominent representative.

In this paper we used the PLS, performed by SmartPLS (Version 3), because of its less stringent distributional assumptions for the variables and error terms and its ability to work with both reflective and formative measurement models. PLS-SEM is widely used for group comparison, investigating the possible presence of a group-effect in the definition of the LVs. The analysis of the invariance of the measures across different groups is necessary when using PLS-SEM for group comparison.

SmartPLS provides permutation-based confidence intervals that allow determining if the correlation between the composite scores of the two groups is significantly lower than one. If the null hypothesis is not rejected, the composite does not differ much in both groups and, therefore, there is compositional invariance. In the next step, permutation-based confidence intervals for the mean values and the variances allow assessing if the composites' mean values and variances differ across groups.

PLS-SEM allows the user to apply three structural model weighting schemes, namely centroid weighting scheme, factor weighting scheme, and path weighting scheme. We chose to use the latter since this weighting scheme is known to provide higher R^2 values for endogenous latent variables and is generally applicable for all kinds of PLS path model specifications and estimations. Moreover, it is not recommended to use the centroid weighting scheme when the path model includes higher-order constructs. The PLS-SEM algorithm stops when the change in the outer weights between two consecutive iterations is smaller than a positive small constant value or the maximum number of iterations is reached. We set 300 as maximum number of iterations, and 0.1^7 as positive small constant for the convergence.

4 Results

Of the original sample of 826 participants, 774 anonymous self-report surveys were returned, resulting in a response rate of 93.7%. Table 1 shows the characteristics of the study sample: the gender ratio of respondents is 0.84, the mean age of the study sample is 39.27 years, in 18–90 age range (standard deviation 12.25); most of them are high school graduates or have a post graduate degree, especially teachers and physicians. The vast majority (89%) has been driving for more than 5 years and 54% of the sample drove a car; only the 27.6% of the interviewed drove both a car and a motorcycle.

The respondents' knowledge about mobile phone use while driving is presented in Table 2. More than 75% of the sample thought that using a mobile phone while driving was one of the main causes of road accidents and was aware that using a hand-free device or headset reduced the risk of road accidents but 28% of them was unaware that this practice was forbidden by law and involved a fine. Moreover, 51% of the sample knew that driver's reaction times while using an electronic device were extended by 50%. Analyzing the "neither agree nor disagree" category data, it emerges that 52.7% of the sample was not aware of highway traffic accidents statistics; 49% was neither agree nor disagree about the claim regarding that using a mobile phone while driving causes a drop in attention level similar to having a blood alcohol level equal to 0.8 g/l. Interestingly, 30% of the sample did not know whether the use of speakerphones while driving entailed any penalties.

In Table 3, attitudes towards the use of mobile phone while driving are shown. Most of the participants thought that mobile phone usage was essential and more than 50% thought that it was necessary for business. According to the 83% of the sample, it was appropriate to use a headset while driving and 47% of them thought that

Table 1 Study sample characteristics

Study population	*N*	Percentage (%)
Gender		
Male	354	45.74
Female	420	54.26
Age		
<30	219	28.29
31–35	133	17.18
36–40	90	11.63
41–45	87	11.24
46–50	111	14.34
>51	134	17.31
Vehicle driven		
Car	418	54.01
Motorcycle	142	18.35
Car and motorcycle	214	27.65
Education		
Primary school	19	2.45
Middle school	55	7.11
High school	378	48.84
Degree	322	41.60
Profession		
Lawyer	24	3.10
Architect	18	2.33
Engineer	30	3.88
Medicine doctor	192	24.81
Employee	72	9.30
Business consultant	18	2.33
Teacher	84	10.85
Dealer	24	3.10
Business owner	18	2.33
Worker	42	5.43
Student	66	8.53
Others	186	24.03
Years of driving		
Some months	18	2.33
1–2 years	30	3.88
3–4 years	37	4.78
>5 years	689	89.02
Smoke		
Smoker	292	37.73
Ex-smoker	124	16.02
Non-smoker	358	46.25

Table 2 Knowledge of respondents toward the use of mobile phone. Neither refers to "Neither agree nor disagree". All values are percentages (%)

N.	Question (Variable)	Agree	Neither	Disagree
K1	Mobile phone use while driving is the main cause of road crashes	79.07	17.83	3.10
K2	Speeding is the main cause of road crashes	63.57	29.46	6.98
K3	Mobile phone related accidents are more frequent on highways	24.81	52.71	22.48
K4	Mobile phone related accidents are more frequent on urban roads	64.34	28.68	6.98
K5	Using hands-free devices reduces the risk of road crashes	75.19	18.60	6.20
K6	Using hands-free devices does not entail penalties	65.12	30.23	4.65
K7	Reading a message takes an average of eight seconds	37.98	46.51	15.50
K8	Reading a message while driving at a speed of 50 km/h is like traveling 111 meters without watching the road	44.96	48.84	6.20
K9	Drivers using a mobile phone have their reaction time extended by 50	51.16	44.19	4.65
K10	Mobile phone use while driving reduces focus as a having a blood alcohol content of 0.8 g/l	40.31	49.61	10.08
K11	Using a mobile phone while driving entails driver license withdrawal	60.47	19.38	20.16
K12	Driving while using a mobile phone entails a fine and driver license points reduction	71.32	18.60	10.08

Table 3 Attitude of respondents toward the use of mobile phone. Neither refers to "Neither agree nor disagree". All values are percentages (%)

N.	Question (Variable)	Agree	Neither	Disagree
A1	Mobile phone use is nowadays necessary	71.32	19.38	9.30
A2	Mobile phone use is indispensable for my line of work	56.59	18.60	24.81
A3	Mobile use while in traffic alleviates wait times	33.33	16.28	50.39
A4	Using earphones while driving is appropriate	82.95	6.98	10.08
A5	Using earphones while driving is bothersome	13.95	20.93	65.12
A6	Using earphones prevents mobile phone use related diseases	47.29	32.56	20.16
A7	In order to reach a destination, asking for directions is more effective	16.28	17.38	65.89
A8	As an auto vehicle add-on, a navigator is fundamental	28.68	29.46	41.86
A9	Do you think mobile phones should be switched off while driving?	25.58	15.50	58.91
A10	Do you think mobile phone use while driving should be allowed by law?	10.08	8.53	81.40
A11	Do you think speed limits should be raised?	18.60	13.18	68.22
A12	Do you think mobile phone use regulation while driving is restrictive?	13.95	19.38	66.67

earphones prevented the onset of mobile phone use related diseases. Interestingly, more than 50% of the disagrees with the possibility to raise speed limits did not consider restrictive the actual regulations regarding the use of the mobile phone while driving. Furthermore, 65.8% of the respondents would never ask directions to a passer-by, although only 28.6% thought that the navigator in a car is a first choice optional.

Behaviors of respondents are listed in Table 4: interestingly, 31% stated they never used a mobile phone while driving, whereas 38% declared that they usually answered a phone call and only 21% stopped the car to answer. Regarding text messages, 24% of the sample admitted to reading them, while only 16% wrote them, of which roughly half aged 30 or less. Only 5% used a mobile phone in order to check their emails while driving and unsurprisingly only 7% switched off the phone while driving. Respondents mainly sought information about the risks concerning mobile phone use while driving, but only 30% kept themselves up to date on laws that regulated its use while driving.

PLS-SEM was performed to formalize a scheme for the interpretation of driving Behavior and to detect its drivers. Starting from the considerations elaborated in the previous sections, we hypothesized that Knowledge and Attitude were exogenous LVs, while Behavior was an endogenous LV. Taking into account the criteria summarized in [41] and considering the choice of the measurement model for latent variables, we adopted a formative model specification for the exogenous factors because it was more plausible to assume each measurement model as an index rather than as a scale. Moreover for each measurement model we observed that each variable was not interchangeable, in fact if we removed a variable from the measurement model we would define different the conceptual domain of the latent variable. For the endogenous factor Behaviour, we chose a reflective measurement model, in fact observing the nature of the factor, we noted that it autonomously exists and all variables relating to it share a common theme. Moreover, the causality flows from the latent factor to the variables and the latter are interchangeability, as in reflective models.

The evaluation model of PLS-PM took place in two steps: the assessment of the outer and inner models. The process therefore started with assessment of the measurement models. In formative models, a first examination regarded the construct validity of formative indicators through theoretic rationale and/or expert opinion [40]. Then, from empirical point of view, the correlation matrix among factors (Table 5) was considered for the evaluation of discriminant validity. Finally, to check the empirically convergent validity of the estimated indicator weighs linking the variables to the corresponding factor, we verified the weights' magnitude and the bootstrapping results for assessing the statistical significance.

Another measurement issue that researchers need to check in formative measurement models is collinearity. The presence of highly correlated variables makes the estimation of their weights in formative models difficult and might result in imprecise values for these weights. In order to check the degree of multicollinearity among the formative indicators the variance inflation factor (VIF, [42]) were computed

Table 4 Behavior of respondents toward the use of mobile phone. All values are percentages (%)

N.	Question (Variable)	Always	Often	Sometimes	Never
B1	Do you ever use a mobile phone while driving?	14.73	7.75	46.51	31.01
B2	Do you use a mobile phone exclusively for work while driving?	2.33	7.75	38.76	51.16
B3	Do you ever answer phone calls while driving?	24.03	13.95	37.21	24.81
B4	Do you ever start phone calls while driving?	21.71	12.40	29.46	36.43
B5	Do you ever read text messages while driving?	14.73	9.60	24.81	51.16
B6	Do you ever send text messages while driving?	10.85	5.43	17.05	66.67
B7	Do you ever read your emails while driving?	3.10	2.33	12.40	82.17
B8	Do you turn off your mobile phone while driving?	4.65	2.33	9.30	83.72
B9	If you receive a phone call or a text while driving, do you stop to answer?	21.71	13.18	31.01	34.11
B10	Do you use earphones while driving?	31.01	20.93	21.71	26.36
B11	Have you ever smoked while driving?	19.38	9.30	17.05	54.26
B12	Have you ever been in a crash because you were using a mobile phone while driving?	3.88	1.55	4.65	89.92
B13	Have you ever been fined for using a mobile phone while driving?	4.65	2.33	3.88	89.15
B14	Do you ever seek information about the risks of mobile phone use while driving?	40.31	13.18	24.81	21.71
B15	Do you keep yourself up to date on laws that regulate mobile phone usage while driving?	36.43	14.73	27.91	20.93

Table 5 Correlation among factors

	Attitude	Behaviour	Knowledge
Attitude	1.00	0.23	0.65
Behaviour	0.23	1.00	0.11
Knowledge	0.65	0.12	1.00

(Table 6). Multicollinearity did not seem to pose a problem, the maximum VIF came to 1.31, which is far below the common cut-off threshold of 10.

In the reflective model, the manifest variables have to be strongly correlated. We verified this in Behavior case. In detail, since reflective indicators have positive inter-correlations, we used the Cronbach's alpha ($0.87 > 0.70$), the average variance extracted ($0.46 > 0.45$) and internal consistency ($0.89 > 0.80$) to empirically

Table 6 Variance inflation factor for the formative indicators

Factor	VIF	Factor	VIF	Factor	VIF
A1	1.18	A7	1.12	K3	1.17
A11	1.15	A8	1.06	K4	1.17
A12	1.13	A9	1.11	K5	1.13
A3	1.18	K10	1.24	K6	1.18
A4	1.06	K12	1.01	K8	1.15
A5	1.07	K2	1.11	K9	1.31
A6	1.03				

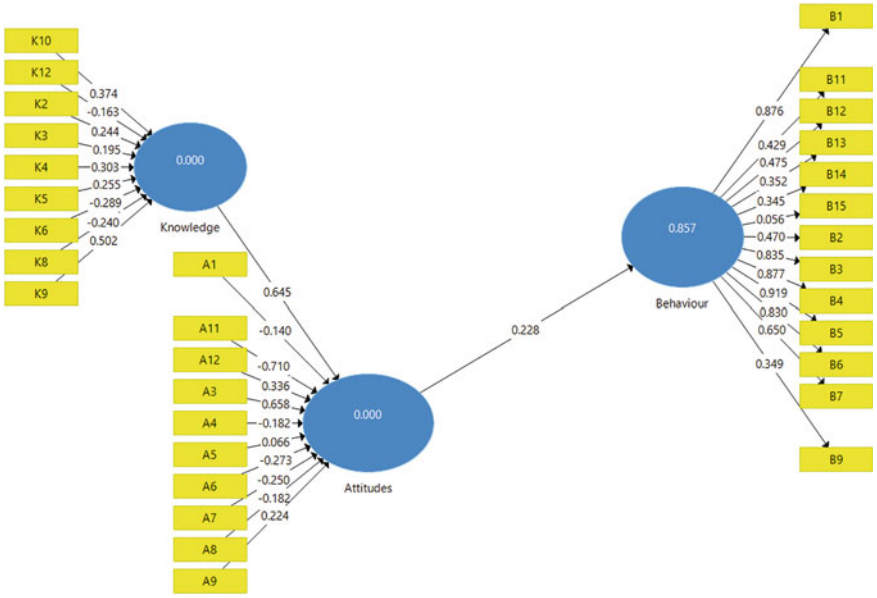


Fig. 1 Structural equation model—path analysis

assess the individual and composite reliability of the indicators. All these measures confirmed the suitableness of the reflective measurement model.

The PLS estimations showed that the relationships between Behavior and Attitude and Attitude and Knowledge were statistically significant (Fig. 1).

The goodness of the model was ultimately very strong ($R^2 = 0.856$). With regards to the path coefficients, we observed that the impact of the Knowledge on the Attitude was considerably greater (0.645) than the impact of the Attitude on the Behavior (0.228). Also the indirect impact of the Knowledge on the Behavior resulted being important (0.147). It is noteworthy that these three impacts and all the outer loadings for latent variables were statistically significant. The direct effect of Knowledge on Behavior was tested but this effect eventually resulted not statistically significant.

5 Discussion

This survey reports detailed information on knowledge, attitudes, and behavior regarding the use of mobile phones while driving in a metropolitan area. Our findings show that there is no correlation between mobile phone use while driving and gender of the interviewed [56], which is contrary to some studies that underscore gender as a differentiating factor: in the case of text messaging while driving, indeed, recent findings define male drivers as more likely to text than female ones [23]. According to our data, 22% of the sample reported using the phone while driving, a smaller percentage compared to the findings from other international studies: Ukraine (34%; [23]), Spain (60%; [19]) and Australia (43%; [55]). To the question related to the frequency of mobile phone usage while driving, 31% of our sample answered "never"; this percentage is significantly higher if compared to 2% of the interviewed by an North-American study stating that they never texted while driving under any circumstances [12].

Our data is consistent with previous observational studies performed in the US, where 31.4% of drivers answered the phone while driving, whereas 16.6% sent text messages or made phone calls [25]; our data report that 38% of the respondents usually answered the phone, while 16% admitted to texting while driving. The least frequent behavior while driving was therefore sending text messages, which was confirmed in other international studies [23, 32]; this suggests that drivers consider writing a text message as leading to a higher level of distraction. This is consistent with previous research on mobile phone distracted driving [47], but it is divergent from another study in which 47% of adults and more than 50% of teens admitted to text messaging while driving [27]. Our findings show an high perception of risk of accidents due to cellphone use while driving in the sample, which is demonstrated by the fact that the vast majority (79.07%) regards it as the main cause of traffic crashes; this is consistent with results from a recent study [9], while being at odds with older findings [56, 57]. Notwithstanding their knowledge on the matter, our interviewed widely agreed that mobile phone use while driving is almost a necessity nowadays: considering that about half of our sample population is composed by post graduates, this data agrees with previous findings that show that the higher the education level, the more likely the interviewed find using a phone while driving necessary, despite the awareness of the danger it entails. The explanation to this phenomenon might be found in the fact that they value their time more than individuals with lesser education [11]. This kind of belief suggests that the perceived benefits of using a mobile phone while driving outweigh the risk of this behavior, as previously observed [56]. According to data collected, more than 50% of respondents thought that mobile phones use while driving was necessary for business, but only a low percentage of them used it solely for work purposes (10%), unlike other studies in which it was found that 75% of phone calls were work-related [31] and in which it was found that drivers tend to use mobile phones more for business than for personal reasons [17, 54].

In the sample interviewed, the use of the headset was widespread, and since 47% thought that using earphones prevented diseases related to mobile phone use, it seemed that our sample was much more concerned by the health risks due to the disproportionate use of the mobile phone than by the risk of road accidents: in fact, only 40% of them declared that they constantly sought information about the risks of cell phone use while driving. Headsets were used by 52% of drivers, in agreement to the findings from a recent study in which drivers reported that the use of a hands-free device is safer than the use of a handheld phone [23]. About 30% of respondents stated that they never used a mobile phone while driving, and its use seemed to be more related to the need to answer phone calls, rather than to the need of reading or writing text messages. It emerged that most of the sample would not ask passers-by for directions, but only 28% believed that the navigator is one of the main accessories to equip a car with; this could be a sign that one of smartphones' most frequently used function while driving was the navigator app. Amongst drivers reporting mobile phone use while driving, 3.88% was involved in a road accident while using the phone, a much lower percentage compared to other studies: in fact in the US mobile phone distraction alone explained about 25% of road crashes [29, 37].

5.1 Limitations

The study was limited by the survey capturing only self-reported behaviors and the drivers may have felt pressure to provide socially acceptable answers; however social desirability bias may have been somewhat allayed since the participants were assured of anonymity and confidentiality. The survey also did not include some activities potentially performed with mobile phones by the younger drivers, such as taking selfies or playing games [38]. Finally, the average age of the studied population was not very high, which could affect the study, since many elders may not own a mobile phone or be able to use all its functions—such as reading email or texting—and as such these additional data might have changed our results.

6 Conclusions

The results of the present research supported the conclusion that the model well represented the collected data according to the result of the goodness-of-fit test. Similarly to earlier studies, this paper confirmed the goodness of the general structural model in helping to understand and explain how Knowledge, Attitude and Behavior are related. The analyzed population showed a good Knowledge on the subject together with positive Attitudes, and there was a general agreement that using a mobile phone while driving is considered unacceptable, even though the employed Behaviors are knowingly inappropriate according to Italian laws. Through our research we dis-

covered that the relatively elevated education level of the sample and the greater driving experience (measured in years of driving license) of the participants were proven as inversely associated with the Behaviors examined; this means that while the experimental results of this survey can be used for the creation of targeted educational programs, community-based interventions and legal regulations, it might be fundamental to act more firmly in order to directly improve people's overall Behavior while driving. All these measures alone, in fact, may not be sufficient to reduce a phenomenon that is so deeply rooted in the population. This ever-growing phenomenon closely follows the technological evolution of our society and it results in an important indicator of how indispensable mobile phones have become in our daily life, a factor being in turn itself dependent on the increasing functions that can be performed through these devices. Considering that - as previously stated - this phenomenon has a strong impact on the increase in road accidents, on the economy and on public health, another solution might be to promote more restrictive regulations establishing a greater number of controls, using not only qualified personnel, but also innovative technologies possibly suitable for detecting real-time hands-on use of the mobile phone while driving.

References

1. European Road Safety Observatory: Annual accident report 2018. Tech. Rep. https://ec.europa.eu/transport/road_safety/system/files/2021-07/asr2018.pdf
2. International Transport Forum: Road safety annual report 2019: Australia. Tech. Rep. https://www.itf-oecd.org/sites/default/files/australia-road-safety.pdf
3. International Transport Forum: Road safety annual report 2019: Italy. Tech. Rep. https://www.itf-oecd.org/sites/default/files/italy-road-safety.pdf
4. Istat Press Release 2019: Road accidents—preliminary estimates: January–June 2019. Tech. Rep. https://www.istat.it/it/files//2019/12/incidenti-stradali2019_stima-gennaio-giugno_EN.pdf
5. The kap survey model (knowledge, attitudes, and practices). Tech. Rep. https://www.medecinsdumonde.org/en/actualites/publications/2012/02/20/kap-survey-model-knowledge-attitude-and-practices
6. National Center for Statistics and Analysis 2019: Driver electronic device use in 2018 (traffic safety facts research note. Report No. dot hs 812 818). National highway traffic safety administration, Washington, DC. Tech. Rep. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812818.pdf
7. Who global status report on road safety 2018. Tech. Rep. https://www.who.int/publications-detail/global-status-report-on-road-safety-2018
8. Adeola, R., Omorogbe, A., Johnson, A.: Get the message: a teen distracted driving program. J. Trauma. Nurs. **23**(6), 312–320 (2016)
9. Al-Jasser, F., Mohamed, A., Choudry, A., Youssef, R.: Mobile phone use while driving and the risk of collision: a study among preparatory year students at king Saud University, Riyadh, Saudi Arabia. J. Fam. Community Med. **25**(2), 102–107 (2018)
10. Alraeesi, F., Farzin, F., Abdouli, K., Sherif, F., Almarzooqi, K., AlAbdool, N.: Smoking behavior, knowledge, attitude, and practice among patients attending primary healthcare clinics in Dubai, United Arab Emirates. J. Family Med. Prim. Care **9**(1), 315–320 (2020)
11. Asensio, J., Matas, A.: Commuters' valuation of travel time variability. Res. Part E Logist. Transp. Rev. **44**(6), 1074–1085 (2008)

12. Atchley, P., Atwood, S., Boulton, A.: The choice to text and drive in younger drivers: behavior may shape attitude. Accid. Anal. Prev. **43**(1), 134–142 (2011)
13. Bianchi, A., Phillips, J.: Psychological predictors of problem mobile phone use. Cyberpsychol. Behav. **8**(1), 39–51 (2005)
14. Consiglio, W., Driscoll, P., Witte, M., Berg, W.: Effect of cellular telephone conversations and other potential interference on reaction time in a braking response. Accid. Anal. Prev **35**(4), 495–500 (2003)
15. De Pretto, L., Acreman, S., Ashfold, M., Mohankumar, S., Campos-Arceiz, A.: The link between knowledge, attitudes and practices in relation to atmospheric haze pollution in peninsular Malaysia. PLoS One **10**(12), e0143655 (2015)
16. Donkor, I., Gyedu, A., Edusei, A., Ebel, B., Donkor, P.: Mobile phone use among commercial drivers in Ghana: an important threat to road safety. Ghana Med. **52**(3), 122–126 (2018)
17. Eost, C., Flyte, M.: An investigation into the use of the car as a mobile office. Appl. Ergon. **29**(5), 383–388 (1998)
18. Gariazzo, C., Stafoggia, M., Bruzzone, S., Pelliccioni, A., Forastiere, F.: Association between mobile phone traffic volume and road crash fatalities: a population-based case-crossover study. Accid. Anal. Prev. **115**, 25–33 (2018)
19. Gras, M., Cunill, M., Sullman, M., Planes, M., Aymerich, M., Font-Mayolas, S.: Mobile phone use while driving in a sample of Spanish university workers. Appl. Ergon. **39**(2), 347–355 (2007)
20. Hancock, P., Lesch, M., Simmons, L., Mouloua, M.: Distraction effects of phone use during a crucial driving maneuver. Accid. Anal. Prev. **35**(4), 501–514 (2003)
21. Harman, H.H.: Modern Factor Analysis. University of Chicago Press, Chicago (1960)
22. Hassen, A., Godesso, A., Abebe, L., Girma, E.: Risky driving behaviors for road traffic accident among drivers in Mekele city, northern Ethiopia. BMC Res. Notes **4**(535) (2003)
23. Hill, T., Sullman, M., Stephens, A.: Mobile phone involvement, beliefs, and texting while driving in Ukraine. Accid. Anal. Prev. **125**, 124–131 (2019)
24. Horn, J.L.: A rationale and test for the number of factors in factor analysis. Psychometrika **30**(2), 179–185 (1965)
25. Huisingh, C., Griffin, R., McGwin, G.: The prevalence of distraction among passenger vehicle drivers: a roadside observational approach. Traffic Inj. Prev. **16**(2), 140–146 (2015)
26. Jöreskog, K.: A general method for estimating a linear structural equation system. ETS Res. Rep. Ser. **2** (1970)
27. Llerena, L., Aronow, K., Macleod, J., Bard, M., Salzman, S., Greene, W., Haider, A., Schupper, A.: An evidence-based review: distracted driver. J. Trauma. Acute Care Surg. **78**(1), 147–152 (2015)
28. Lorini, C., Pellegrino, E., Mannocci, F., Allodi, G., Indiani, L., Mersi, A., Petrioli, G., Santini, M., Garofalo, G., Bonaccorsi, G.: Use of seat belts and mobile phone while driving in florence: trend from 2005 to 2009. Epidemiol. Prev. **36**(1), 34–40 (2012)
29. Mirman, J., Durbin, D., Lee, Y., Seifert, S.: Adolescent and adult drivers' mobile phone use while driving with different interlocutors. Accid. Anal. Prev. **104**, 18–23 (2017)
30. Montuori, P., Triassi, M., Sarnacchiaro, P.: The consumption of genetically modified foods in Italian high school students. Food Qual. Prefer. **26**(2), 246–251 (2012)
31. Musicant, O., Lotan, T., Albert, G.: Do we really need to use our smartphones while driving. Accid. Anal. Prev. **85**, 13–21 (2015)
32. Nemme, H., White, K.: Texting while driving: psychosocial influences on young people's texting intentions and behavior. Accid. Anal. Prev. **42**(4), 1257–1265 (2010)
33. Nevin, P., Blanar, L., Kirk, A., Freedheim, A., Kaufman, R., Hitchcock, L., Maeser, J., Ebel, B.: "i wasn't texting; i was just reading an email ...": a qualitative study of distracted driving enforcement in Washington state. Inj. Prev. **23**(3), 165–170 (2017)
34. Nobile, C., Montuori, P., Diaco, E., Villari, P.: Healthcare personnel and hand decontamination in intensive care units: Knowledge, attitudes, and behaviour in Italy. J. Hosp. Inf. **51**(3), 226–232 (2002)

35. Oviedo-Trespalacios, O., King, M., Haque, M., Washington, S.: Risk factors of mobile phone use while driving in Queensland: prevalence, attitudes, crash risk perception, and task-management strategies. PLoS ONE **12**(9), e0183361 (2017)
36. Oviedo-Trespalacios, O., Scott-Parker, B.: Transcultural validation and reliability of the spanish version of the behaviour of young novice driver scale (bynds) in a Colombian young driver population. Transp. Res. Part F Traff. Psychol. Behav. **49**, 188–204 (2017)
37. Pless, C., Pless, B.: Mobile phones and driving. BMJ **348**, 1193 (2014)
38. Postelnicu, C., Machidon, O., Girbacia, F., Voinea, G., Duguleana, M.: Effects of playing mobile games while driving, distributed, ambient, and pervasive interactions. In: Proceedings of the Fourth International Conference, DAPI 2016, pp. 291–301 (2016)
39. Pöysti, L., Rajalina, S., Summala, H.: Factors influencing the use of cellular (mobile) phone during driving and hazards while using it. Accid. Anal. Prev. **37**(1), 47–51 (2005)
40. Rossiter, J.R.: The coarse procedure for scale development in marketing. Int. J. Res. Marketing **19**, 305–335 (2002)
41. Sarnacchiaro, P., Boccia, F.: Some remarks on measurement models in the structural equation model: an application for socially responsible food consumption. J. Appl. Stat. **45**(7), 1193–1208 (2018)
42. Sen, A., Srivastava, M.: Regression Analysis: Theory, Methods and Applications. Springer, New York, USA (1990)
43. Shaaban, K.: Investigating cell phone use while driving in Qatar. Procedia Soc. Behav. Sci. 104 (2014)
44. Skierkowski, D., Wood, R.: To text or not to text? The importance of text messaging among college-aged youth. Comput. Human Behav. **28**(2), 744–756 (2012)
45. Sullman, M., Baas, P.: Mobile phone use amongst New Zealand drivers. Transp. Res. Part F Traff. Psychol. Behav. **7**(2), 95–105 (2004)
46. Sullman, M., Przepiorka, A., Prat, F., Blachnio, A.: The role of beliefs in the use of hands-free and handheld mobile phones while driving. Comput. Human Behav. **9**, 187–194 (2018)
47. Tison, J., Chaudhary, N., Cosgrove, L.: National phone survey on distracted driving attitudes and behaviors. National Highway Traffic Safety Administration, Washington, DC. Tech. Rep. (2011)
48. Tomer-Fishman, T.: Distraction in driving by the use of electronic communication devices. The Israeli Road Safety Authority (in Hebrew) (2010)
49. Trivedi, N., Haynie, D., Bible, J., Liu, D., Simons-Morton, B.: Cell phone use while driving: prospective association with emerging adult use. Accid. Anal. Prev. **16**, 450–455 (2017)
50. Truong, L., Nguyen, H., De Gruyter, C.: Mobile phone use among motorcyclists and electric bike riders: a case study of Hanoi, Vietnam. Accid. Anal. Prev. **91**, 208–215 (2016)
51. Valent, F., Del Pin, M., Mattiussi, E., Palese, A.: Prevalence of mobile phone use among drivers: direct observation in Udine (northern Italy). Epidemiol. Prev. **44**(2–3), 171–178 (2020)
52. Velicer, W.F.: Determining the number of components from the matrix of partial correlations. Psychometrika **41**, 321–327 (1976)
53. Walsh, S., White, K.: Ring, ring, why did I make that call? Mobile phone beliefs and behavior among Australian university students. Youth Stud. Aust. **25**(3), 49–57 (2006)
54. Walsh, S., White, K., Hyde, M.B.W.: Dialing and driving: factors influencing intentions to use a mobile phone while driving. Accid. Anal. Prev. **40**(6), 1893–1900 (2008)
55. White, K., Hyde, M., Walsh, S., Watson, B.: Mobile phone use while driving: an investigation of the beliefs influencing drivers' hand-free and hand-held mobile phone use. Transp. Res. Part F Traff. Psychol. Behav. **13**(1), 9–20 (2010)
56. White, M., Eiser, J., Harris, P.: Risk perceptions of mobile phone use while driving. Risk Anal. **24**(2), 323–334 (2004)
57. Wogalter, M., Mayhorn, C.: Perceptions of driver distraction by cellular phone users and nonusers. Hum. Factors **47**(2), 455–467 (2005)
58. Wold, H.: Path models with latent variables: the nipals approach. In: Blalock, H., Aganbegian, A., Borodkin, F., Boudon, R., Capecchi, V. (eds.) Quantitative Sociology, International Perspectives on Mathematical and Statistical Modeling, pp. 307–357. Academic Press (1975)

Impact of Emergency Online Classes on Students' Motivation and Engagement in University During the Covid-19 Pandemic: A Study Case

Isabella Morlini

Abstract The Covid-19 pandemic has had dramatic impact on many dimensions of living and studying conditions of students at University. This paper analyses student satisfaction and motivation during the lockdown period and try to understand whether different socio-economic and environmental conditions have influenced needs and demands of students during the emergency online didactics. Drawing from the results of a questionnaire administered to students enrolled in the University of Modena and Reggio Emilia, this research is aimed at describing which factors, beyond the quality and the professionalism of the lecturers and the quality of the education received, influence the satisfaction with the online learning experience and impact on students' motivations and perceived engagement. Moreover, the study investigates the pandemic's direct effects on gender differences and inequalities, analysing the obstacles affecting the self-organization of study at home.

Keywords Gender difference · Multiple correspondence analysis · Online survey · Student satisfaction · k-means cluster analysis

1 Introduction

The health emergency brought about by COVID-19 has produced a radical and rapid change in university life. In March 2020, remote teaching became the rule almost overnight in all Italian universities, with new implications in the landscape of educational learning. Classes, examinations and laboratories were suddenly reorganized by a collective effort that benefited from previous experimentation and innovation in teaching methods like, for example, blended learning (see e.g., [24]). Emergency remote teaching was a temporary shift of instructional delivery to an alternate mode due to crisis circumstances [16] and involved the use of fully remote teaching solutions for instruction and education that would otherwise be delivered face-to-face or as blended courses. For this reason, at the core of the challenge was not only the

I. Morlini (✉)
University of Modena & Reggio Emilia, Modena, Italy
e-mail: isabella.morlini@unimore.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_18

technical and remote delivery of all classes, but especially the array of tools and practices concerning engagement of the students and developing of their attitudes toward online learning, like, for example, project works. As outlined in Conole [7], online learning is both social and cognitive process, not merely a matter of information transmission via remote information technologies. The importance of understanding students' motivation and engagement in an online environment is shown by the significant and recent amount of research on this topic (see, e.g., [2, 11, 12, 18, 20, 27]). Several studies were conducted to determine the factors that are expected to have an effect on students' remote emergency learning experience [1, 14, 19]. The goal of this work is to reach an insight on the needs of the university students from a customer-oriented perspective and to analyse the socioeconomic and environmental determinants of their satisfaction and motivations in the emergency remote teaching system. Using data collected with an online questionnaire administered to students enrolled in the University of Modena & Reggio Emilia and performing cluster and multiple correspondence analysis, we try, particularly, to give an answer to the following research questions:

- Is there an association between students' satisfaction for online experience and intrinsic motivations and attitude to distance learning?
- Do socioeconomic problems influence the distance learning experience?
- Which are the determinants for satisfaction and engagement in online education?
- Do gender and working status play a role in students' satisfaction?
- Do university related characteristics like the area of the course in which the student is enrolled and the course year, influence students' satisfaction?

We also try to investigate the presence of a possible gender gap among students, comparing the differences in indicators related to study organization, concentration, material conditions and obstacles affecting the self-organization of study at home.

While several studies analyze the experiences related to distance learning in sample of students or teachers in school during the first wave of the COVID-19 pandemic in Italy (see, e.g., [6, 13]), the impact on student's learning in Italian Universities remains an aspect largely unexplored. The study conducted by Calandri et al. [4] among Italian university students focuses on the role of concerns, change in peer and family relationships rather than in the analysis of the distance learning experience.

The paper is organized as follows: Sect. 2 briefly illustrates the questionnaire and the sample, Sect. 3 reports the most important results of multiple correspondence analysis and k-means cluster analysis, Sect. 4 deals with the impact of remote teaching on gender inequalities and differences and Sect. 5 gives some concluding remarks, focusing on recommendations for, eventually, future online didactics.

2 Data Collection

The online survey was implemented by means of Survey Monkey, with an individual link sent to 27,792 students during the period April, 8th 2020–May, 2nd 2020. The questionnaire consists of 36 questions grouped into four sections:

1. General information on home trips caused by the emergency, living conditions and ongoing problems and changes.
2. Organization of study with respect to the teaching materials available, the timing and methods of the organization of study.
3. Distance learning, with the focus on attendance and satisfaction, specific difficulties, conditions of concentration and interest, aspects that were missing and those that were appreciated, open questions on strengths and weaknesses of distance learning and suggestions and proposals; information on internships and working conditions.

For a detailed description of the questionnaire and of the respondents we refer to Russo et al. [25]. Data collected, metadata and descriptive statistics are available online at http://dx.doi.org/10.25431/11380_1203517. For some questions the answer is dichotomous (1 = yes, 0 = no) while for other questions the answer is ordinal with four categories, or it is open-ended. In this work, we have considered only closed-questions related to socioeconomic and material problems, student motivation and attitude to online learning, task orientation and engagement. To both reduce the number of parameters in the multivariate models and simplify output results for gaining clear and useful information, we have re-coded all polytomous variables in order to have only binary variables of the type presence/absence, substituting categories "not at all" and "rarely" with the code 0 and categories "enough" and "very much" with 1. Moreover, we have dichotomized the rate of the satisfaction with the global learning experience re-coding a rate from 1 to 5 into 0 (not satisfied) and a rate from 6 to 10 into 1 (satisfied). Previous multivariate analyses on variables with four modalities have led to results difficult to summarize and therefore less informative. Comparing main findings on original ordinal variables (see [26]) we see that results are insensitive to the dichotomization rule adopted. On the other hand, classification of students by considering the combinations of all material conditions and choices related to studies and life organisation results more complicated to interpret with original ordinal responses.

The average rate of complete answers is 19.2% (5,341 records). The participation rate varies widely by area, year of enrolment, achieved credits and gender (Table 1, last column). Considering the area, students enrolled in a course in the area of Society and Culture show the highest participation, while students enrolled in a course in Science or Technology show the lowest participation. Considering the year, students in the second year shows the highest participation while students in the first year show the lowest. Regarding credits achieved and gender, students with achieved credits above the median of the total credits achieved in the same course and in the same year, participated more than students with credits achieved below the median and female students participated more than males.

Table 1 Frequency distributions of students in the population (enrolled students) and in the sample (completed survey) by area, year of enrolment, achieved credits and gender and response rates

		Enrolled students (%)	Completed survey (%)	Response rate (%)
Total		100	100	19.2
Area	Health	8.76	8.95	19.6
	Life	6.86	5.73	16.1
	Science	6.69	4.91	14.1
	Society and culture	51.09	61.97	23.3
	Technology	25.03	18.37	14.1
	Erasmus	1.58	0.07	0.9
Year of enrolment	1	27.98	22.73	15.6
	2	18.85	23.67	24.1
	3	26.51	24.73	17.9
	4	10.90	12.68	22.3
	5	15.00	15.32	19.6
	6	0.75	0.88	22.6
Achieved credits	Below the median	55.08	48.19	34.4
	Above the median	43.11	51.69	46.1
	N/A	1.81	0.11	1.2
Gender	Female	51.74	65.74	24.4
	Male	46.68	34.17	14.1
	Not responding	1.58	0.09	1.1

For the different areas and years of enrolment, the response percentages in the sample are quite like the percentages of those enrolled students in the population (Table 1), with the exceptions of areas and years with the lowest and the highest respondent rate, mentioned above. Considering the gender, we note an over-representation of females: the percentage in the sample is 65.6% while the percentage of females in the population is 52.5%. Considering the credits, we note an over-representation of students with credits achieved above the median: the percentage in the sample is 43.1% while the percentage in the population is 51.7%. It seems therefore that female students and, in general, students progressing satisfactory in university studies, are the most engaged in challenges produced by emergency remote learning. It is worth noting that also in the survey conducted by Calandri et al. [4] through an online questionnaire on the Limesurvey platform, by means of the nonrandom snowball sampling technique, females result over-represented, amounting to the 83% of the respondents.

3 Results of the Multivariate Analyses

The first two factors of the multiple correspondence analysis [15] performed over the 28 dichotomous variables considered in this work, explain the 25.6% of the total variability, a percentage higher than the threshold $0.95^{28} = 23.8\%$, suggested in Zani and Cerioli [29]. With this threshold, we retain the number of factors accounting, on average, for at least the 95% of the variance of each original variable. Considering that the total variance is monotonically increasing with the number of variables, Zani and Cerioli suggested this threshold for the variance explained criterion [17] depending on the number of variables. The screenplot (Fig. 1) indicates the optimal number of latent factors as one or two. Analysing the principal coordinates (see Table 2) we may evaluate the first factor as an index of high unsatisfaction with the global online experience, strong demotivation and disengagement. On the contrary, the second factor may be considered as an index of engagement and moderate unsatisfaction. We then retain two factors for gaining insights into the impact of remote emergency learning. The first factor loads strongly on variables related to the lack of attitude to online learning (like inability to organize daily studying activities effectively disorientation due to the many channels of communications) and to the lack of task orientation (like having hard time concentrating, inability to extricate among recordings, inability to take good notes, and accumulation of lessons). The second factor loads only on few variables related to demotivation and lack of attitudes and it is not associated with task orientation. Among socioeconomic and material problems, variables loading on both factors are those related to the care of the family and, of course, those related to proper electronic devices and internet connection. Economic problems due to the pandemic seem not to be related to unsatisfaction and amotivation. Results show that technological, beside human support, is one of the causes affecting students' dissatisfaction of remote teaching. Bad network services and technical issues still play a critical role in an online teaching environment. Another important cause is the feeling of the students as if they were studying all the time, causing stress and difficulties in focusing on the lectures. This emphasizes the importance of a proper and flexible organization of the courses online.

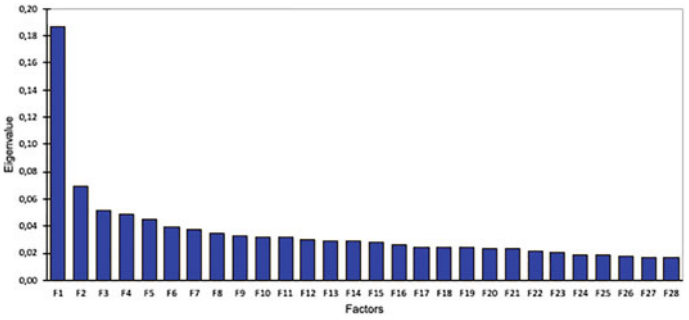


Fig. 1 Scree plot of the multiple correspondence analysis

Table 2 Principal coordinates of the multiple correspondence analysis

	Variable number	Questions	F1	F2
	1	Satisfied with the global distance learning experience: NO	**0.811**	0.163
		Satisfied with the global distance learning experience: YES	**−0.320**	−0.064
Socioeconomic and material problems	2	Economic problems due to the pandemic: NO	−0.066	−0.107
		Economic problems due to the pandemic: YES	0.272	**0.441**
	3	Problems due to the condition of family members or friends: NO	−0.045	−0.051
		Problems due to the condition of family members or friends: YES	**0.475**	**0.541**
	4	No time to study due to taking care of family or live-in friends: NO	−0.069	−0.107
		No time to study due to taking care of family or live-in friends: YES	**0.586**	**0.912**
	5	No proper internet connection: NO	−0.090	−0.055
		No proper internet connection: YES	**0.539**	**0.330**
	6	No proper electron device for studying activities: NO	−0.031	−0.027
		No proper electron device for studying activities: YES	**0.731**	**0.636**
Motivation	7	Progressing equally on all subjects in the semester: NO	0.127	**0.237**
		Progressing equally on all subjects in the semester: YES	**−0.344**	**−0.642**
	8	Appreciating the chance of keeping abreast of all subjects more easily: NO	**0.483**	0.107
		Appreciating the chance of keeping abreast of all subjects more easily: YES	**−0.504**	−0.111

(continued)

Table 2 (continued)

	Variable number	Questions	F1	F2
	9	Appreciating the possibility to pause the recordings and listening again: NO	**0.616**	**0.399**
		Appreciating the possibility to pause the recordings and listening again: Yes	−0.113	−0.074
	10	Appreciating the interactions among students through many channels: NO	0.094	−0.004
		Appreciating the interactions among students through many channels: YES	**−0.359**	0.015
	11	Appreciating having the resources of the course at all time everywhere: NO	**0.578**	0.246
		Appreciating having the resources of the course at all time everywhere: YES	−0.214	−0.091
Lack of task orientation	12	NOT understanding how to organize the studying activities: NO	**−0.465**	−0.200
		NOT understanding how to organize the studying activities: YES	**0.575**	0.247
	13	Accumulating lessons create difficulties: NO	**−0.620**	−0.144
		Accumulating lessons create difficulties: YES	**0.475**	0.110
	14	Don't know how to extricate among the different recordings: NO	**−0.390**	−0.130
		Don't know how to extricate among the different recordings: YES	**0.781**	0.260
	15	Having a hard time concentrating: NO	**−0.797**	−0.194
		Having a hard time concentrating: YES	**0.456**	0.111
	16	Lacking the will to study: NO	**−0.538**	−0.265
		Lacking the will to study: YES	**0.425**	0.210

(continued)

Table 2 (continued)

	Variable number	Questions	F1	F2
	17	Difficulties in following the lessons and tacking good notes: NO	**−0.385**	−0.050
		Difficulties in following the lessons and tacking good notes: YES	**0.772**	0.099
Lack of engagement	18	The absence of involvement makes it difficult to stay focused: NO	**−0.626**	**0.378**
		The absence of involvement makes it difficult to stay focused: YES	**0.510**	**−0.308**
	19	The absence of interaction does not allow for enough explanations: NO	**−0.412**	**0.394**
		The absence of interaction does not allow for enough explanations: YES	**0.441**	**−0.422**
	20	Lacking the possibility to ask for explanations: NO	**−0.339**	**0.429**
		Lacking the possibility to ask for explanations: YES	**0.442**	**−0.560**
	21	Lacking the chance to interact face to face: NO	**−0.716**	**0.880**
		Lacking the chance to interact face to face: YES	**0.221**	**−0.271**
	22	Lacking the stimuli given in class: NO	**−0.848**	**0.554**
		Lacking the stimuli given in class: Yes	**0.407**	**−0.266**
Lack of attitude to online learning	23	It fatigues spending lot of time in front of the screen: NO	**−0.672**	0.295
		It fatigues spending lot of time in front of the screen: YES	**0.306**	−0.134
	24	It takes longer to follow a recording lesson: NO	**−0.647**	**0.389**
		It takes longer to follow a recording lesson: YES	**0.322**	−0.193
	25	Distracted by other family care or needs at home: NO	**−0.389**	−0.267

(continued)

Table 2 (continued)

	Variable number	Questions	F1	F2
		Distracted by other family care or needs at home: YES	**0.419**	0.288
	26	Many channels of communications disorient: NO	**−0.190**	−0.108
		Many channels of communications disorient: YES	**0.650**	**0.370**
	27	Don't able to organize daily studying activities effectively: NO	**−0.648**	**−0.357**
		Don't able to organize daily studying activities effectively: YES	**0.558**	**0.307**
	28	Lacking the weekly schedule of lessons: NO	**−0.819**	**0.402**
		Lacking the weekly schedule of lessons: YES	**0.408**	−0.200

Positive aspects like the ability to view the recorded lectures and the course materials online and offline and interact with many channels, seem to be less important than negative aspects like lower participation and involvement in online environments than in traditional environments and the difficulty in asking questions.

Considering all previous variables except for the satisfaction with the global learning experience, we perform a k-means cluster analysis with Euclidean distance. Analysing the decomposition of the deviance within and between groups obtained, we choose a partition into 4 clusters as the best partition for the trade-off between number of groups and homogeneity inside the groups. Table 3 reports the central coordinates in each cluster. Figure 2 reports the biplot of the multiple correspondence analysis.

The multiple correspondence analysis is performed on group membership, satisfaction and independent variables like gender, working status, course year, area of the course in which the student is enrolled, credits acquired (above or below the median of credits acquired by students in the same course and in the same year) and the presence of economic problem aside from the pandemic.

Analysing the central coordinates of the groups and the position of clusters in the biplot, with respect to the modalities "not globally satisfied with online learning" and "globally satisfied with online learning" we see that Cluster 1 identifies the group of students (31.3% of the total) totally unsatisfied. These students are only characterized by lack of attitude, lack of engagement and orientation. Cluster 2 is the group of the globally satisfied students (21.6% of the total), mostly enrolled in the third year of a

Table 3 Central coordinates of the 4 clusters: in rows the clusters and in the columns the variables

Cl	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	0	0	0	0	0	1	0	1	1	1	1	1
2	0	0	0	0	0	0	1	1	0	1	0	0	0	0
3	0	0	0	0	0	0	1	1	0	1	0	0	0	0
4	0	0	0	0	0	0	1	1	0	1	1	1	0	1

Cl	16	17	18	19	20	21	22	23	24	25	26	27	28
1	1	1	1	1	1	1	1	1	1	1	0	1	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	1	1	1	1	1	1	0	0	0	1
4	1	0	0	0	0	1	1	1	0	1	0	1	1

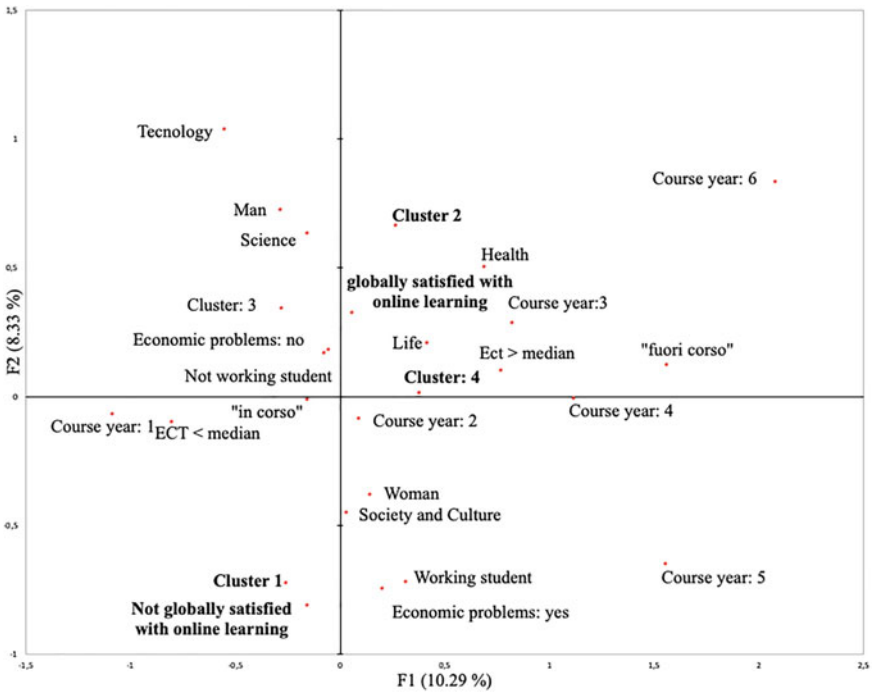


Fig. 2 Biplot of the multiple correspondence analysis

course in Health or Life and characterized by attitude to online learning, task orientation and self-engagement. Motivation seems not to discriminate between satisfied and unsatisfied students. Clusters 3 (24.1% of the total) and 4 (23% of the total) identify students with intermediate level of engagement and satisfaction. Socioeconomic problems due to the pandemic do not discriminate groups. However, economic

problems aside from the pandemic do impact negatively on the student's satisfaction. The analysis also reveals that, beyond student's aptitude to online learning, the area of course plays an important role in satisfaction with online learning. Courses in the area of society and culture seem to be unsuitable for remote teaching. As outlined in Szeto [28] positive aspects like having accessible and varied means of communication and quick responses between student and lecturers, are important in some specific teaching and learning contexts, such as technological courses. In these courses, students are more likely to appreciate the flexibility in emailing the lecturers. Another finding is that students enrolled in the first year are more likely to be unsatisfied and disengaged: indeed, for these students, the emergency online teaching was the first experience with online learning, while the other students in the University of Modena and Reggio Emilia have had the chance to attend some blended courses. For what concern the working status, surprisingly working students are more likely to be dissatisfied than not working students. On the contrary, the status "in corso" or "fuori corso", peculiar of the Italian universities, is not a determinant of student's satisfaction.

4 A Study of the Gender Differences and Inequalities

A large amount of literature has analyzed the impact of Covid-19 on the work conditions of woman and men academics during the lockdown, paying attention to the comparisons of the unequal effects on the research activities (see, e.g., [3, 8, 9, 22, 23]). Recurring finding in this literature is that during the pandemic many academics, and especially women academics, have been less productive, because they found a difficult task working at home and, especially, doing research. Less attention has been devoted to possible gender inequalities in academic students. In this section, the questions explored is whether the changes in the conditions of studies due to the pandemic negatively affected more female than men. Table 4 reports variables for which the percentages of responses "Yes" and "No" is significantly different between men and women (p-value < 0.001). It is meaningful that the percentage of females not having time to study at home in the emergency period due to taking care of family or live-in friends (variable 4) is significantly higher than the percentage of males. It seems that nonacademic responsibilities such as those related to care and housework are unequal distributed in households even for students. The significantly higher percentage of women than the percentage of men distracted by family care or needs at home (variable 25) supports this gender inequality among students. Economic problems due to the pandemic also affects more female students than males (variable 2).

Even though both men and women pointed out the challenges connected with online teaching, we observe gender differences regarding the specificity of these challenges and the different emotional response. Analysing variables 10, 18 and 19 reported in Table 4, we note that women are more capable of reach online interactions and involvements with lecturers and students. Indeed, women appreciating

Table 4 Variables for which the percentage of respondents Yes and NO is significantly different (p-value < 0.001) between men and women

Variable		% of respondents	
		Female (%)	Male (%)
2	Economic problems due to the pandemic: NO	83.56	87.09
	Economic problems due to the pandemic: YES	**16.44**	**12.91**
4	No time to study due to taking care of family or live-in friends: NO	88.12	92.07
	No time to study due to taking care of family or live-in friends: YES	**11.88**	**7.06**
10	Appreciating the interactions among students through many channels: NO	77.62	82.50
	Appreciating the interactions among students through many channels: YES	**22.38**	**17.50**
14	Don't know how to extricate among the different recordings: NO	63.62	72.48
	Don't know how to extricate among the different recordings: YES	**36.38**	**27.52**
18	The absence of involvement makes it difficult to stay focused: NO	46.71	41.79
	The absence of involvement makes it difficult to stay focused: YES	**53.29**	**58.21**
19	The absence of interaction does not allow for enough explanations: NO	53.40	48.58
	The absence of interaction does not allow for enough explanations: YES	**46.60**	**51.42**
23	It fatigues spending lot of time in front of the screen: NO	25.97	41.58
	It fatigues spending lot of time in front of the screen: YES	**74.03**	**58.42**
24	It takes longer to follow a recording lesson: NO	31.11	37.25
	It takes longer to follow a recording lesson: YES	**68.89**	**62.75**
25	Distracted by other family care or needs at home: NO	49.33	56.78
	Distracted by other family care or needs at home: YES	**50.67**	43.22

the possibility of interacting through many channels with other students are significantly more numerous than men and, conversely, men not able to stay focused due to absence of involvement and not able to get enough explanations due to absence of interactions, are significantly more numerous than women.

On the other hand, responses in variables 23 and 24 suggests that female students perceive online teaching more time consuming and demanding than face-to-face classes, compared to men. They also seem to suffer more technological obstacles like the difficulty in extricating among the different recordings.

We may be more conservative and confident about the significance of the results by considering the Bonferroni correction for multiple comparisons [10, 21]. With this correction, percentages of female and male respondents remain significantly different for a family wise error rate $\alpha = 0.001$ (since p-values are less than 0.001/28) for variables 4, 10, 14, 23, 24, 25. For variables 2, 18, 19, the difference of percentages of responses remains significant considering a family wise error rate $\alpha = 0.05$ (since p-values are less than 0.05/28). Performing multiple correspondent analysis separately for men and women, we reach very similar results in terms of principal coordinates of the first two factors. Similarly, performing cluster analysis separately for men and women, we reach identical central coordinates for the partitions into 4 clusters. These results show that, aside the specific variables listed in Table 4, there is not gender difference in the latent structure of motivation, orientation, engagement and attitude to online learning.

5 Conclusions

This study highlights the experience and the effects of the emergency remote teaching environment on students of the University of Modena and Reggio Emilia, in Italy, during the lockdown period. A strength of the study is the large sample size (5,341 students) in which each scientific area and each course year is well represented. A main limitation is, of course, the self-selection bias and the limited percentage of respondents with respect to the total population to which the questionnaire was sent (this small percentage may have also been caused by the restricted number of emails of reminder sent during the survey campaign). Given that the characteristics of the population of the University students are well-known, the bias due to the limited percentages of respondents could have been partially reduced by proceeding with a quota sampling. However, the self-selection bias due to the fact that the compilation of questionnaire was not mandatory and required time and accuracy, could not have been weakened by sampling strategies. Indeed, as outlined in Sect. 2, female students and, in general, students progressing satisfactory in university studies, resulted the most engaged in compiling the questionnaire. The limited generalizability of the results, however, should not detract from the valuable findings. The different analyses show that there is a strong association between student satisfaction for online experience and attitude to distance learning, engagement and task orientation. Student intrinsic motivations seem to play a less relevant role. Socioeconomic and material

problems due to the pandemic influence the distance learning experience but are not a determinant for satisfaction, engagement and motivation. On the contrary, economic problems aside from the pandemic are associated with dissatisfaction, as long as the working status. The area of the course in which the student is enrolled, and the course year influence the satisfaction. On the other hand, the number of credits acquired and the status "in corso" or "fuori corso" (peculiar of the Italian university system) are not linked to motivation ad self-engagement.

The study confirms that valid predictors of student perception and engagement are, among other factors related to task orientation and attitude, the ability to organize the daily activities effectively, the ability to stay focused and not be distracted, the ability to extricate among the recorded lessons and not to be disoriented by the different channels of communications. Students without attitudes toward online learning are incapable of recognising positive aspects like the possibility to pause the recordings and listening again, the possibility of having all the resources of the course at all time and everywhere and the possibility of keeping abreast of all subjects. Results suggest that attitudes and tasks toward online can be gained across time: students enrolled in the third year of courses that are more experienced with university courses and have probably experimented some blended learning before the emergency period, are more likely to be satisfied with remote didactics and are more likely to be engaged. These results and some other insights of this study, like the critical role still played by bad internet services and technical issues in an online teaching environment, can be extended beyond the pandemic, when considering substituting traditional face to face learning with remote teaching.

Regarding the study of a possible gender gap induced by the pandemic, the analysis of the main challenges generated by the emergency remote teaching for female students and males reveals that the home environment is not the same for men and women in terms of the ability to stay focused on their intellectual work. Indeed, female students are more likely to be distracted by family care or needs than males. In addition, female students claim more than men economic problems due to the pandemic. Regarding the differences in facing the new challenges, responses in the different questions suggest that female students perceive online teaching more time consuming and demanding than face-to-face classes, compared to men and that they also seem to suffer more technological obstacles like the difficulty in extricating among the different recordings. On the other hand, women seem more capable of reach online interactions and involvements with lecturers and students. Similar results for teachers in Italian universities are reported in [5]: the Italian survey on the university teachers' perspectives and their emotional conditions conducted by the authors shows significant impairments in sleep patterns and loss of energy, with female teachers having greater difficulty concentrating than their male colleagues.

References

1. Almossa, S.Y.: University students' perspectives toward learning and assessment during COVID-19. Educ. Inf. Technol. (2021). https://doi.org/10.1007/s10639-021-10554-8
2. Bolliger, D.U., Halupa, C.: Online student perceptions of engagement, transactional distance and outcomes. Distance Educ. **39**(3), 299–316 (2018)
3. Boncori, I.: The never-ending shift: a feminist reflection on living and organizing academic lives during the coronavirus pandemic. Gend. Work Organ. **27**(5), 677–682 (2020)
4. Calandri, E., Graziano, F., Begotti, T., Cattelino, E., Gattino, S., Rollero, C., Fedi, A.: Adjustment to Covid-19 lockdown among Italian university students: the role of concerns, change in peer and family relationships and in learning skills, emotional and academic self-efficacy on depressive symptoms. Front. Psychol. **12** (2021). https://doi.org/10.3389/fpsyg.2021.643088
5. Casacchia, M., Cifone, M.G., Giusti, L., Fabiani, L., Gatto, R., Lancia, L., Cinque, B., Petrucci, C., Giannoni, M., Ippoliti, R., Frattaroli, A.R., Macchiarelli, G., Roncone, R.: Distance education during COVID 19: an Italian survey on the university teachers' perspectives and their emotional conditions. BMC Med. Educ. 21–335 (2021)
6. Commodari, E., La Rosa, V.L.: Adolescents and distance learning during the first wave of the Covid 19 pandemic in Italy: what impact on student's well-being and learning processes and future prospects? Investig. Health Psychol. Educ. **11**, 726–735 (2021)
7. Conole, G.: Learning Design in Practice: Fostering Different Pedagogical Approaches. Taylor & Francis, Milton Keynes (2021)
8. Corbera, E., Anguelovski, I., Honey-Rosés, J., Ruiz-Mallén, I.: Academia in the time of COVID-19: towards an ethics of care. Plan. Theory Pract. **21**(2), 191–199 (2020)
9. Cui, R., Ding, H., Zhu, F.: Gender inequality in research productivity during the COVID-19 pandemic. SSRN Electron. J. 1–30 (2020)
10. Dunn, O.G.: Multiple comparison among means. J. Am. Stat. Assoc. **56**(293), 52–64 (1961)
11. Ellis, R., Bliuc, A.: Exploring new elements of the student approaches to learning framework: the role of online learning technologies in student learning. Act. Learn. High. Educ. **20**(1), 11–24 (2019)
12. Ferrer, J., Ringer, A., Saville, K., Parris, M.A., Kashi, K.: Student's motivation and engagement on higher education: the importance of attitude to online learning. High. Educ. (2020). https://doi.org/10.1007/s10734-020-00657-5
13. Giovannella, C., Passarelli, M., Persico, D.: The effects of the Covid-19 pandemic on Italian learning ecosystems: the schoolteacher's perspective at the steady state. Interact. Des. Archit. J. **45**, 264–286 (2020)
14. Gopal, R., Singh, V., Aggarwal, A.: Impact of online classes on the satisfaction and performance of students during the pandemic period of COVID 19. Educ. Inf. Technol. (2021). https://doi.org/10.1007/s10639-021-10523-1
15. Greenacre, M., Blasius, J.: Multiple Correspondence Analysis and Related Methods. Chapman and Hall, NY (2006)
16. Hodges, C., Moore, S., Lockee, B., Trust, T., Bond, A.: The Difference Between Emergency Remote Teaching and Online Learning. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning (2020)
17. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002)
18. Kahu, E.R., Nelson, K.: Student engagement in the educational interface: understanding the mechanisms of student success. High. Educ. Res. Dev. **37**(1), 58–71 (2018)
19. Mailizar, M., Burg, D., Maulina, S.: Examining university students' behavioural intention to use e-learning during the COVID-19 pandemic: an extended TAM model. Educ. Inf. Technol. (2021). https://doi.org/10.1007/s10639-021-10557-5
20. Martin, F., Bolliger, D.H.: Engagement matters: student perceptions on the importance of engagement strategies in the online learning environment. Online Learn. **22**(1), 205–222 (2018)
21. Miller, L.G.: Simultaneous Statistical Inference. Springer (1966)
22. Minello, A., Martucci, S., Manzo, L.: The pandemic and the academic mothers: present hardships and future perspectives. Eur. Soc. 1–13 (2020)

23. Nash, M., Churchill, B.: Caring during COVID-19: a gendered analysis of Australian university responses to managing remote working and caring responsibilities. Gend. Work Organ. **27**(5), 833–846 (2020)
24. Purnomo, A., Kurniawan, B., Aristin, N.: Motivation to learn independently through blended learning. Adv. Soc. Sci. Educ. Humanit. Res. **330**, 261–264 (2019)
25. Russo, M., Alboni, F., Colombini, S., Morlini, I., Pavone, P., Sartori, L.: Covid-19 e Studenti Unimore: come l'emergenza cambia lo studio e l'esperienza universitaria. Demb Working Paper Series, n. 173 (2020)
26. Russo, M., Alboni, F., Colombini, S., Morlini, I., Pavone, P., Sartori, L.: Learning online: remote teaching and university student's engagement. Italian J. Appl. Stat. (2022)
27. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices and future directions. Contemp Educ Psychol **61** (2020)
28. Szeto, E.: Community of Inquiry as an instructional approach: what effects of teaching, social and cognitive presences are there in blended synchronous learning and teaching? Comput. Educ. **81**, 191–201 (2015)
29. Zani, S., Cerioli, A.: Analisi dei Dati e Data Mining per le Decisioni Aziendali, Giuffrè, Milano (2007)

Local Heterogeneities in Population Growth and Decline. A Spatial Analysis of Italian Municipalities

Federico Benassi, Annalisa Busetta, Gerardo Gallo, and Manuela Stranges

Abstract Spatially unequal demographic dynamics lead to a progressive fragility of a territory and its socio-economic system. In Italy, municipalities characterized by demographic malaise tend to be increasingly small in size and peripheral in location, and their local spatial aggregation increased over time. A spatial approach is here proposed to investigate the dynamics across time and space of the population variations in Italian municipalities from 1981 to 2011. Global and local spatial autocorrelation analysis and several models of regression were run using as study variable the average growth rates at municipality level. The spatial autocorrelation of the study variable is quite high and stable over time. The regression results show that spatial models (SAM and SAR) outperform the non-spatial model (OLS) and that SAR is the best model. The results also underline that the variation of population is significantly affected by its values in the neighbouring municipalities, confirming the spatial nature of the phenomenon. The presence of schools in the municipality emerges as a key factor for the increase/decrease of the population. Moreover, the decomposition of the effects into direct and indirect effects shows that all the independent variables produce their effects almost 70% directly and 30% indirectly.

Keywords Demographic malaise · Italy · Spatial regression models · Spatial demography · Local analysis

A. Busetta (✉)
Department of Economics, Business and Statistics (SEAS), University of Palermo, Viale delle scienze, ed. 13, 90141 Palermo, Italy
e-mail: annalisa.busetta@unipa.it

F. Benassi
Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, 80133 Naples, Italy

G. Gallo
Italian National Institute of Statistics (Istat), P.zza Guglielmo Marconi, 26/c, 00144 Rome, Italy

M. Stranges
Department of Economics, Statistics and Finance "Giovanni Anania" (DESF), University of Calabria, Ponte Pietro Bucci, Cubo 0/C, 87036 Arcavacata di Rende, CS, Italy

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_19

1 Introduction and Brief Literature Review

Not a single southern or western European Union member state recorded a population decline during 1990–2015 (neither Italy nor Germany who had a negative natural balance in the period). However, according to Eurostat projections, European Union will reach its maximum population in 2026 (with 449.3 million) and then will experience a new phase of population decline, which will cause a shrinking of the population size in the subsequent decades (up to 441.2 million in 2050 and to 416.1 million in 2100).

Demographic studies agree that in an ageing society characterized by a low or sometimes a lowest-low fertility rate, migration has become the driver of demographic change [37]. From a demographic point of view, a population decline leads to a movement of people to more prosperous areas in a downward spiral in which the most schooled and most active individuals emigrate from the declining areas, while the poorest and most dependent ones remain. Indeed, migration has negative impacts on the population age structure of the sending areas (acceleration of population ageing, decrease in the size and ageing of labour force) and on its dynamic (reduction in number of births, relative increase in mortality), but can also mitigate the population decline in many receiving areas. In this line, some studies have focused on intra-regional migration trends and estimated its effect on out-flows [37].

Some studies noticed that population decline is not uniform across territories, but focused on the analysis of significant differences in the rate and direction of population change [42]. Golini et al. [24] studying the Italian situation at the municipality level showed that the "more dynamic" areas are contrasted by others characterized by a demographic malaise. A large strain of the scientific literature studies depopulation in the context of the urban–rural divide. Indeed, even if most of the studies associate depopulation with rural areas [28, 29, 38], some studies emphasize that the dividing line between growing and shrinking regions does not follow only an urban–rural divide [35].

As the number of developed countries and regions across the world reporting a population decline is growing, over the past decade both scholars and public policymakers have become increasingly interested in this issue [19]. Among others, the studies of Coleman and Rowthorn [17] and that of Rees et al. [39] debate whether population decline is good or bad for countries and territories. The environmental literature emphasizes the pros of population decline (reduction of environmental impact, greater sustainability, lowering of traffic congestion and its consequences, release of more housing space, …), while the economic one mainly focuses on its negative consequences (shrinking of economic activities, regional markets and investment, decline of innovation and downward spiral of economic activity, …).

Since the patterns of population growth and decline are not uniform over territories (with strong differences within the same countries in which some areas lose population while others gain), there is an unequal demographic dynamics that could lead to

a progressive fragility of the territory and its socio-economic system,[1] with consequences in terms of sustainability, development and well-being. Indeed "unbalanced" territorial models of demographic development are a risk for the socio-economic system, increasing its exposure to natural risks and environmental issues, social conflict, limiting the possibilities of overall growth and causing a worsening of the quality of life [15, 25]. Territories subject to depopulation are realities that are gradually becoming weaker and unsafe. Progressive abandonment, marginalization and neglect often result in greater exposure to exogenous shocks, such as harmful and uncontrollable natural events or lasting economic crises [18, 30]. On the other hand, territories that grow very quickly—typically large urban and metropolitan areas—have other problems that arise from the processes of concentration and, in some cases, the real spatial saturation of the population. These territories, which are also considered the engines of economic growth, as well as the main actors in the processes of globalization [22, 43, 44], are often characterized by a lower quality of life, by high rates of pollution and by a growing social conflict due to typically urban phenomena, such as residential segregation of particular subgroups of the population, marginality and social deviance, as well as extreme poverty [16, 27, 36].

To overcome the consequences of heterogeneous demographic dynamics, at the European level, the need to support the development of "polycentric territorial systems" has been recognized, or rather territorial areas characterized by well-interconnected medium-sized cities [21]. The European Commission affirms that a territorial redistribution of the population and a balanced growth of the territories are necessary conditions for a significant, lasting and sustainable development of the various local realities [21, 51]. In this perspective, strategies and policies to deal with the problems associated with shrinking regions—i.e. isolated and/or peripheral realities that are in systematic demographic decline—have been developed both at the European [20] and the Italian level.[2]

Based on these premises, the present study deals with the spatial dimension of the demographic growth and decline of Italian municipalities. This paper is structured as follows. In Sect. 2 we discuss the Italian case, presenting some descriptive results about the average annual growth rate and some reflections about its impact on the age structure in the local contexts. In Sect. 3 we describe the data and methods used for the empirical analysis. Section 4 is devoted to the presentation of the results of a spatial autocorrelation and the dynamics of population growth and decline at the local level over a long period (1981–2019). In Sect. 5 we present and discuss the results of OLS, SEM and SAR models, through which we individualize the underlying factors that affected the variation of the average population growth rates in the period 2011–2019. Finally, a brief discussion of the main results is given in Sect. 6.

[1] Each territory (country, region, province, municipality etc.) that is administrative and geographically identified (i.e. has boundaries) defines automatically a socio-economic system which is composed by the people and the firms that live and act in that "space".

[2] "Strategia Nazionale Aree Interne" https://www.agenziacoesione.gov.it/strategia-nazionale-aree-interne/.

2 The Italian Case

As for many other European countries, in Italy the population trend is strongly territorially differentiated with some municipalities that show a systematic loss of population and others with an equally continuous and significant increase [8, 9, 42]. Previous studies of population at the municiple level in Italy showed that there are some "more dynamic" situations contrasted by others characterized by demographic malaise that tend to be increasingly small in size and become peripheral in location [24, 42]. This pattern is evident by analysing the maps of Italian municipalities by average annual growth rate (Fig. 1), which show a worsening of the situation in the last decade with an increasing number of municipalities in slight or even intense demographic decline (around 50% of the municipalities belong to these two categories) and a decline of municipalities in intense demographic growth (in 2011–2019 about 6% belong to this category).

From the maps of Fig. 1, the relevance of space in defining the temporal dynamics of the population growth at local level seems quite clear [11, 45]. These preliminary descriptive results motivated us to analyse the Italian situation using more refined methods.

A spatially unbalanced demographic growth naturally has repercussions on the age structure of the population, which is, again, strongly unequal from municipality to municipality, reproducing the patterns just observed (Fig. 1). As it is evident from the maps of Fig. 2, aging is a spatially heterogeneous phenomenon, characterized by specific geographies that hint at direct relationships with depopulation [40, 41].

The territories where the population is most aged and where the gap between young and old is, at least quantitatively, extremely significant are, in 2019, mostly the areas that had been subject to systematic depopulation in the period 1981–2019 (Fig. 1). Thus, a real vicious circle of demographic marginality can be observed: local contexts that decline demographically tend to be those that age the most. Given the relationships between age structure and productivity/wealth, well highlighted by some authors [7], these local contexts are destined to become—if only because of the different age structure—less productive and less wealthy. In fact, the attitudes, behaviors, and preferences of individuals vary with age and, therefore, the age structure of the population can greatly affect the economic performance of a given territorial context. Obviously, this is not just an economic issue, given that territories that are more aged will also have different needs in relation, above all, to the care of the elderly and the need for health structures and hospitals. Similar is the situation regarding the demographic dependence index, i.e., the ratio between the population of non-working age (young and old) and that of working age which—theoretically—should produce income. Also in this case, territorial inequalities are evident and significant. All of this, translated onto a local scale, implies that areas with a declining population and characterized by a comparatively more unbalanced age structure of the population will become progressively less competitive and therefore unable to retain their own resident population—especially those of working age and youth—and attract new ones.

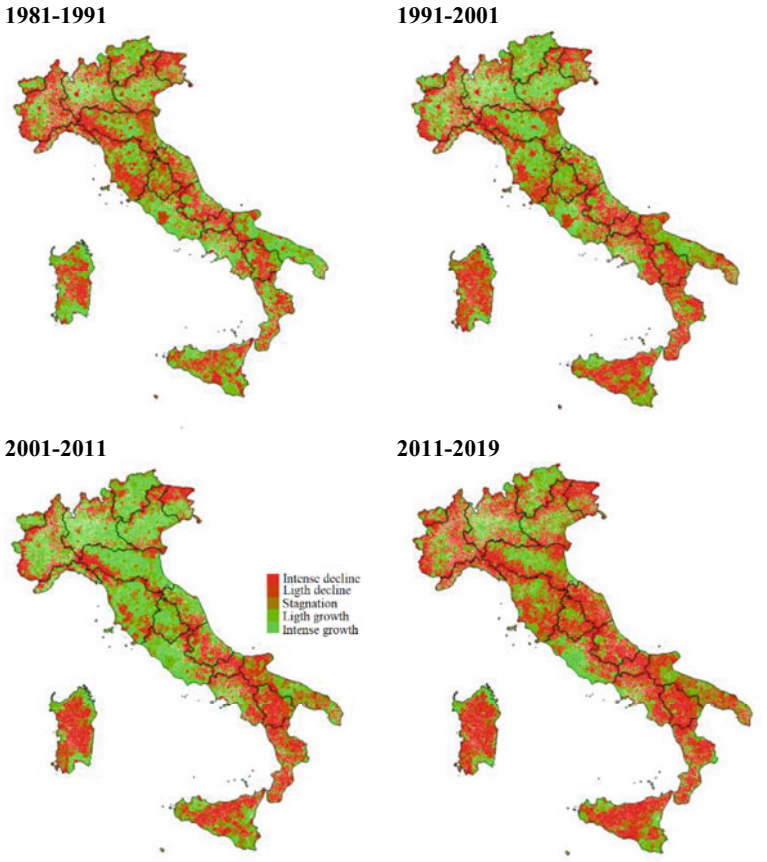


Fig. 1 Average annual growth rates (per thousand)[(a)]. [(a)]The class are defined as follows: intense decline (<−8.0‰); slight decline (−8.0‰ ≤ −2.0‰); stagnation (>−2.0‰ ≤ 2.0‰); slight growth (>2.0‰ ≤ 8.0‰); intense growth (>8.0‰). *Source* Our elaboration on Census data 1981–2011. Data on 2019 stem from population registers (pre-census correction)

3 Data and Methods

On the basis of the literature review and the descriptive findings in the previous paragraph, this paper tries to investigate the local demographic dynamics that occurred in Italy in the last 40 years following a spatial approach of analysis. The study is based upon data on population provided by Istat (Italian National Institute of Statistics), coming from the demographic censuses (1981–2011) and from population registers at 2019 (pre-census correction). The statistical units of analysis are the 7,926 Italian municipalities. Their number and geographies have been reconstructed to a fixed time (01.01.2019) and maintained stable across time so that we can make accurate comparisons. Geographical data on municipalities (shape files) are provided by Istat.

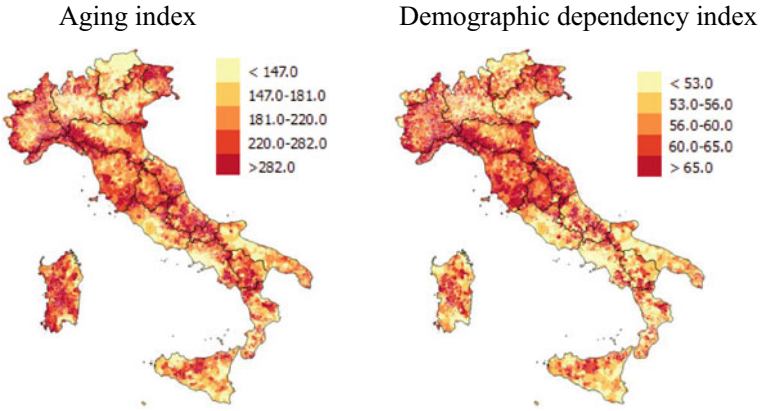


Fig. 2 Aging index[a] and Demographic dependency index[b]. Italian municipalities. 2019. [a]The index is obtained as a percentage ratio between population aged 65 and over and population 0–14 years old. [b]The index is obtained as percentage ratio between non active aged population (65 and over y.o. and 0–14 y.o.) and the active age population (15–64 y.o.). The classes of the maps are quintiles of the reference indexes. *Source* Our elaboration on data from population registers (pre-census correction)

The choice of using the municipalities as the unit of analysis (instead of NUTS-2, regions; or NUTS-3, provinces) is driven by the fact that, as shown in the maps of Fig. 1, there are substantial sub-regional and sub-provincial heterogeneities in the spatial pattern of growth or decline, with some areas of municipalities having a growing trend and some others a declining trend. The focus on Italian municipalities has been used in other demographic papers, such as Caltabiano et al. [14], which explores the organisation of family systems, finding significant differences at the municipality level that are mainly undetected when larger territorial units of analysis are considered. More recently, Salvati et al. [45] used the same territorial scale for detecting the spatial dimension of total fertility rate and crude birth rate in the period 2002–2018 for Italy.

In all the spatial analyses here proposed, the spatial weight matrix (W) is obtained as "Queen" contiguity matrix of the first order: two municipalities are neighbours if and only they geographically share a boundary or a vertex. It should be noted that the number of statistical units on which the spatial analysis is carried out became 7,912 because of the elimination of 14 neighbourless municipalities. The statistical analysis is composed of two parts. In the first one (Sect. 4), the attention is devoted to measuring the global and local spatial autocorrelation of the average annual growth rates in the periods 1981–1991, 1991–2001, 2001–2011, and 2011–2019 using the global Moran *I* index and its local version.[3] The second part (Sect. 5) focuses on the estimation of aspatial (OLS) and spatial (SEM and SAR) regression models.[4]

[3] This part is carried out with GeoDa (version 1.18 10.12.2020) by Luc Anselin.

[4] Regression models (aspatial and spatial) are estimated by using R Studio [33]. Thematic maps, including the ones of Fig. 1, are created using Qgis "Odense" version 3.30.2. For methodological aspects of regression models here adopted, refer to Golgher and Voss [23].

The dependent variable is the average annual growth rates for the last period (2011–2019) while the covariates are demographic and socio-economic indicators stemming from the 2011 Census. In this section, we aim to understand how different demo-socio-economic dimensions directly affect the demographic growth and decline of Italian municipalities. We therefore run a regression analysis in which the 2011–2019 average annual growth rate is related to a set of covariates measured in the 2011 Census. In doing that we cannot ignore the spatial nature of the phenomena here observed and partially studied in the previous section. The regression analysis is therefore based on the comparison between three models: Ordinary Least Square (OLS), Spatial Error Model (SEM) and Spatial Lag Model (SAR). The first model is the classic linear standard regression model, whereas SAR and SEM are spatial econometric models [4, 31].

Following Sun et al. [47] and Golgher and Voss [23], a SAR model is a model that examines how the dependent variable (y) is influenced by the value assumed by the same variable in adjacent spatial units (in our case, municipalities). The general SAR model is defined as follows:

$$y = \rho W y + X\beta + \varepsilon$$

where the spatial lag parameter (ρ) refers to the estimate of how the average dependent variable in neighbouring spatial units (municipalities) is associated with the same variable for a local spatial unit (municipality). Then W is a matrix of spatial weights, X is a matrix of independent variables, β is a vector of the regression coefficients and ε is a vector of residuals.

By contrast, an SEM estimates the extent to which the OLS residual of a Municipality is correlated with that in its adjacent municipalities. The model is defined by the following equation:

$$y = X\beta + u$$

$$u = \lambda \mathrm{W}u + \varepsilon$$

where y is the dependent variable; X is a matrix of the independent variables; β is a vector of the regression coefficients; λ is the error spatial coefficient parameter; ε is a vector of residuals, and Wu is spatial weighting matrix error. The spatial error parameter (λ) measures the strength of the relationship between the average residuals/errors in neighbouring municipalities and the residual/error of a given municipality [47].

One crucial aspect of SAR, and in general of all spatial autoregressive regression models, is that the coefficients cannot be interpreted as in an OLS model, but rather it is necessary to refer to direct and indirect (spatial spillovers) effects [23]. The direct effect *"represents the expected average change across all observations for the dependent variable in a particular region due to an increase of one unit for a specific explanatory variable in this region"* ([23]: 185), while the indirect effect *"represents*

the changes in the dependent variable of a particular region arising from a one-unit increase in an explanatory variable in another region" ([23]: 185).

As already explained, in the analysis we are going to compare a classic regression model (OLS) and the two spatial models, SEM and SAR, using the AIC [1] to determine if one model is better than the other, assuming that the model with the smaller AIC value should be preferred as it is more likely to minimize the information loss in contrast to the true model that generates the observed data [12, 50]. From an interpretative point of view, the comparison between OLS and spatial regression models is straightforward. The OLS model parameters are estimated under the explicit assumption that the observations are independent, that means that changes in values for one observation do not "spill over" to affect values of another observation [23]. Spatial regression models, in contrast, assume that the observations are not independent and that they can exert a reciprocal influence, as the first law of geography of Tobler [48] clearly states. The independent variables used in the model analyse whether the signs of future population growth/decrease can be identified among the characteristics detected for each municipality at a previous point of time, namely the 2011 Census (Table 1).

To avoid redundancy in the information included in the models and manage high correlation between independent and explanatory variables, we reduced the number of covariates so that in all the models the multicollinearity condition number is lower than 30 [6]. The covariates included in the model belong to four dimensions (Table 1): The *demographic dimension* (percentage of preschool children, percentage of elderly over 75, percentage of foreign population), the *socio-economic dimension* and *mobility* (percentage of young people living alone, female activity rates, employment rate of young people aged 15–29 and mobility for study and work reasons);

Table 1 Conceptual dimensions and related covariates used in the regression analysis

Conceptual dimensions	Variables	Sources
Demographic	Percentage of less than 6 years old (%) Percentage of over 75 years old (%) Percentage of foreign people (%)	Population and Housing Census
Socio-economic and mobility	Percentage of youth living alone (%) Female activity rate (%) Youth (15–29 years old) employment rate (%)	Population and Housing Census
Schooling	Presence/absence of primary school	Ministry of Education and Scientific Research
Economic-productive/environment	Share of employees in agricultural sector (%) Share of employees in industrial sector (%)	Population and Housing Census

the *schooling dimension* (presence/absence of primary school) and the *economic-productive and environment dimension* (share of employees in the agricultural and industrial sector).

4 Global and Local Spatial Autocorrelation of Population Growth and Decline

In Fig. 3 we observe the values of the global Moran's *I* index of spatial autocorrelation [34] and the maps related to its local version [3]. The increasing level of spatial clustering of the variable 'average annual growth rate' in Italy is shown globally by the value of Moran's *I*: the index increases from 0.401 in the period 1981–1991, reaches 0.536 in 2001–2011 and then declines to 0.436 in 2011–2019. The positive signs of *I* prove that there is a positive global spatial autocorrelation, that is to say that similar (positive or negative) values of the observed variable (average growth rates) tend to be spatially clustered. Based on the local version of Moran's *I* index each municipality was classified as: (i) High-High (HH) hot spots (high growth rates with similar values among neighbouring municipalities), (ii) Low-Low (LL) cold spots (low growth rates with similar values among neighbouring municipalities), (iii) High-Low (HL) potential spatial outliers (high growth rates with low growth rates among neighbouring municipalities), (iv) Low–High (LH) potential spatial outliers (low values with high growth rates among neighbouring municipalities) or finally (v) units spatially uncorrelated with neighbours (i.e. when the spatial distribution of the observed variable is random).[5]

From the analysis of the four maps in Fig. 3 we note, on the one hand, the transition phases of Italian metropolitan areas from suburbanization (LH clusters) to reurbanization (HH clusters) and, on the other, the consolidation of some demographically 'weak' areas (LL clusters) in the inland areas of the South and in the peripheral areas of the North and the Centre. The final result is a broken up territory in which growing and spatially compact areas are in certain ways opposed to large areas in population decrease. This can be read as a sign of alarm because, as we already underlined, a spatial framework in which some local realities win and others lose represents a sort of dual space that is detrimental to social cohesion and sustainable development [21]. In this perspective it can be useful to compare the nearest time period (2011–2019) and the farthest one (1981–1991). In doing so we can appreciate how the geographical distance between the different clusters of municipalities and, in particular, between hot (High-High clusters) and cold (Low-Low clusters) spots has increased, coming to define 'different spaces' in a rather defined way. In 1981–1991, the first period of observation, the red lumps affected very specific and delineated areas of Italy in correspondence with the large urban and metropolitan centres of both the Centre-North and the South, identifying urban expansion. At the

[5] It is important to keep in mind that the reference to high and low is relative to the mean of the variable, and should not be interpreted in an absolute sense.

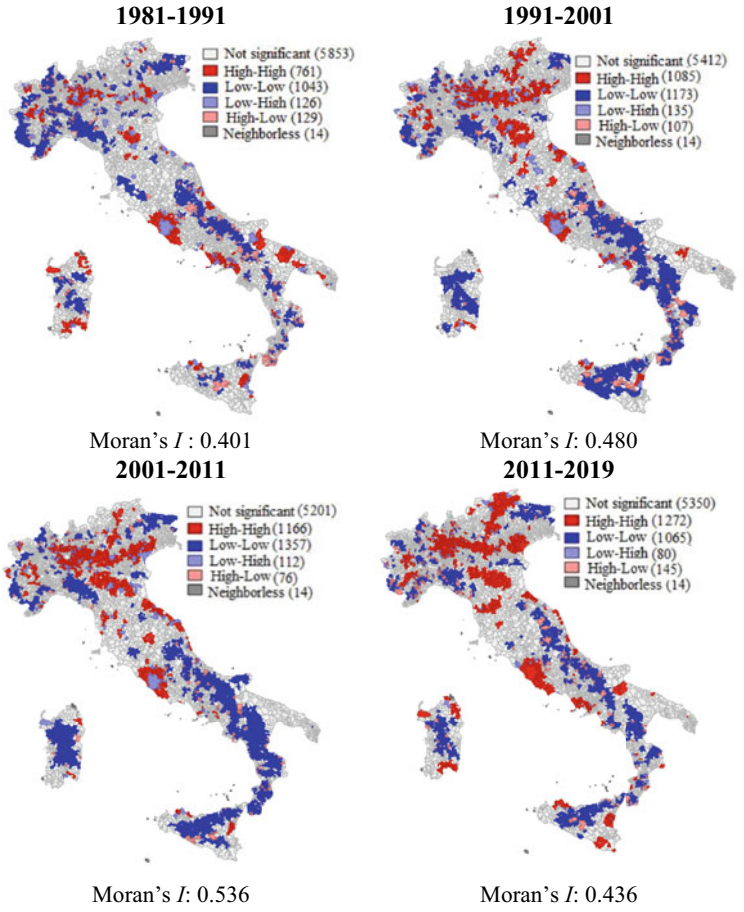


Fig. 3 Global and local indexes of spatial autocorrelation for the variable average annual growth rates (per thousand)[(a)]. [(a)]Statistically significant at $p \leq 0.05$. The number of municipalities belonging to each cluster is reported in the legend among brackets. *Source* Our elaboration on Census data 1981–2011. Data on 2019 are from population registers (pre-census correction)

same time, the blue areas (cold spots) were also spatially and geographically very well circumscribed to some Apennine areas, and inland Sicily and Sardinia, the pre-Alps and some areas of Liguria and Emilia Romagna. In the last period (2011–2019) it is evident how the urban and metropolitan spatial plots have expanded especially in the Centre-North: Lombardy and Veneto are linked by clusters of contiguous hot spots from Milan almost to Venice, also affecting the mountain areas of Trentino South Tyrol, not dissimilar is what can be seen along the territory of the Tuscan Valdarno in a trajectory that connects Pisa to Florence and which, interrupted by the Tuscan-Emilian Apennines, resumes in Emilia connecting, along the east–west route, Bologna with the other provincial capitals. The metropolitan area of Turin is

also quite evident. Rome, the Italian capital, is at the centre of a group of municipalities characterized by "red tiles" that extends from lower Lazio, along the Pontine countryside, and then stops before resuming along the Tyrrhenian coast of Campania with Naples and its hinterland. The major conurbations of Apulia, Sicily and Sardinia are also evident. The cold spots are distributed more widely but always according to precise geographical logics: the internal territories of Sardinia and those of Sicily, the Apennine ridge, internal areas of Liguria and Emilia near the Apennines, alpine border areas. It is well known that the presence of foreign people played an important role in some of these demographic dynamics, and especially in the growth of urban and metropolitan areas, which have grown considerably in the last twenty years and, in particular, for the central municipalities of metropolitan cities and for the municipalities adjacent to them [46].

We can summarize the data of Fig. 3 as shown in Fig. 4. The trend in the number of municipalities with positive (HH or LL) and negative spatial (HL or LH) autocorrelation shows that 2011–2019 represents a period of change with a decrease of municipalities that registered low growth rates with similar or different values among neighbouring municipalities (LL and LH clusters). The same period registered a change in the potential outlier municipalities with an increase of municipalities with high growth rates that have low growth rates among neighbouring municipalities (HL cluster), and a decrease of municipalities with low growth rates that have high growth rates among neighbouring municipalities (LH cluster).

The existence of such a demographic spatial dynamics poses interesting questions about the drivers of population growth and decline. In the next section we model the variation of population that occurred in each municipality in the last period of analysis (2011–2019) and considering additional information about each territorial unit in the analysis.

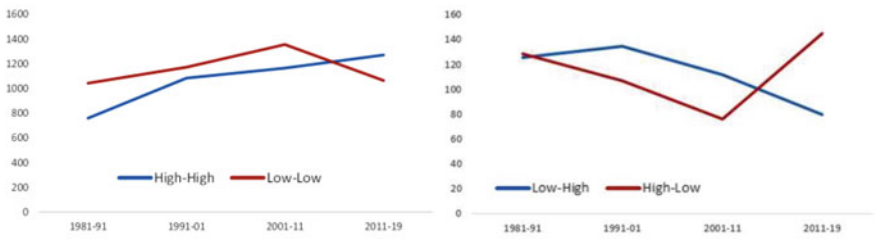


Fig. 4 Number of municipalities with positive (left) and negative (right) spatial auto-correlation. *Source* Our elaboration on Census data 1981–2011. Data on 2019 are from population registers (pre-census correction)

5 Results

Before showing the results of our empirical analysis, we have run some diagnostic tests to compare the performance of the three models. Results of these diagnostics (Table 2) show that the best model is the SAR (or spatial lag) one, which is the one which has the lowest AIC. This is coherent with the results of the diagnostic tests for spatial dependency. In particular, the Robust LM (lag) statistics is significant ($p < 0.05$), while the Robust LM error is not ($p > 0.05$), so according to Anselin's decision tree [5] we should use a spatial lag model.

In Table 3 we report the results of all the three models we have estimated but, for the sake of brevity, we comment only the estimates of the SAR model which—as just shown from the previous diagnostics—has proved to be the best model among the others. Overall, the results of the three are consistent with each other in sign and magnitude.

The results we obtain for the estimated coefficient of the spatial lag of the annual average growth rate (Rho) in the SAR model is equal to 0.31, which indicates positive correlation between the annual average growth rate in a municipality and the average growth rate in its neighbouring municipalities. This means that the variation of y is significantly influenced by the value it assumes in the spatial neighbourhood of each municipality. This confirms the spatial nature of the phenomenon, supporting the hypothesis about the existence of a spatial diffusion process which manifests itself net of the influence of the other explanatory variables included in the model.

About the covariates, the model allows the decomposition of the coefficients into direct (DE) and indirect effects (IE), i.e. the effect of the municipality's level on itself and on the other municipalities. For all the covariates, it turns out that the direct contribution to the average annual growth rate of the municipality is around 70% of the total effect, whereas the effect on the neighbouring municipalities is around 30%. The demographic dimension is significantly and consistently associated with population variation. In particular, the age structure is among the main forerunners of population growth/decrease. The percentage of children up to 6 years is strongly and positively associated with the 2011–2019 change in the population, suggesting a demographic increase, while a high percentage of elderly over 75 years shows a negative association. As reported in Table 3 the municipality direct effect of a 1 percentage

Table 2 Robust diagnostic for spatial dependence, LM multiplier

Test	Values	Probability
Moran's I (error)	19.33	0.00
LM (lag)	551.24	0.00
Robust LM (lag)	183.20	0.00
LM (error)	368.77	0.00
Robust LM (error)	0.73	0.39

Source Our elaboration on Istat data, 2001 and 2011 Census and 2019 population registers (pre census correction)

Table 3 OLS, SEM and SAR models results (dependent variable: average annual growth rate in the period 2011–2019)[(a)]

Variables	OLS	SEM	SAR			
	Coeff	Coeff	Coeff	DE	IE	TE
Intercept	−4.33	−4.67	−3.27			
% less than 6 y.o	1.09	0.92	0.86	0.88	0.38	1.26
% over 75 y.o	−0.78	−0.75	−0.61	−0.62	−0.27	−0.89
% foreign people	0.02	0.01	0.01	0.01	0.00	0.01
% youth living alone	0.09	0.11	0.10	0.10	0.04	0.14
Female activity rate	0.18	0.18	0.16	0.16	0.07	0.23
Youth employment rate	0.04	0.04	0.03	0.03	0.01	0.04
% workers in agriculture	−0.15	−0.14	−0.12	−0.12	−0.05	−0.17
% workers in industry	−0.15	−0.13	−0.11	−0.11	−0.05	−0.16
Study to work mobility	−0.02	−0.02	−0.03	−0.03	−0.01	−0.04
Primary school	1.03	1.11	0.75	0.76	0.33	1.09
λ (spatial error parameter)		0.33				
ρ (spatial lag parameter)			0.31			
AIC	51386.7	51039.7	50871.6			

[(a)]All the coefficients are statistically significant at $p < 0.05$. Decomposition of effect is obtained by a simulation procedure

Source Our elaboration on Istat data, 2001 and 2011 Census and 2019 population registers (pre census correction)

point increase in the percentage of children in the municipality increases the average annual growth rate by 0.88% points. The across-municipality spillover effect of a 1 percentage point increase in the share of children increases the average annual growth rate by 0.38% points on average. The percentage of foreigners, although statistically significant, makes little contribution to the growth/decrease of the population: the own-municipality direct effect of a 1 percentage point increase in the percentage of foreigners increases the average annual growth rate by 0.0096% points. The across-municipality spillover effect of a 1 percentage point increase in the percentage of foreigners increases the average annual growth rate by 0.0041% points on average.

The socio-economic dimension also plays an interesting but small role. The share of young people living alone is significantly associated with an increase in the population of the last 10 years. Leaving the family of origin is therefore confirmed as an important phase of transition that constitutes the necessary premise for the formation of the family (and the other stages of transition to the adulthood) and an important driving force for stimulating population growth through fertility. At the municipal level, the percentage of youths living alone in 2011 is positively associated with the variation of the population in the following years with a direct effect of 0.10 and a spillover effect of 0.04. The employment status of more vulnerable groups (women and young people) is particularly interesting for giving information not only about

the labour market, but also on the economic autonomy necessary for the transition to adulthood and parenthood. Although studies still do not completely agree about the sign of the relationship between female labour force participation and fertility at the micro and macro levels, all agree that economic factors matter ([2, 13, 26, 49]. Busetta and Giambalvo [13] showed that an increase in female participation in the labour market (macro) at the regional level is associated with a small but significant increase in the probability of having a first child, whereas women's participation in the labour market (micro) is negatively associated with it. Innocenti et al. [26] find a clear positive association between the "Economic Complexity" (measured considering many different economic indicators) and fertility change across Italian provinces between 2006 and 2015, net of traditional fertility predictors. From our models it results that the municipality's direct effect of a 1 percentage point increase in the female activity rate increases the average annual growth rate of the same municipality in the next ten years by 0.16% points and the average annual growth rate of the neighbouring municipalities by 0.07% points on average. Lower, even if smaller, the effect of the youth employment rate that is of 0.03 for the municipality and 0.01 for its neighbours on average. Also the ability to reach places of study and work in a short time is certainly one of the factors that slows down the depopulation of a territory. At the municipal level, the percentage of the population that moves to reach the places of study and work registered in 2011 is negatively associated with the variation in the following years with a direct effect of −0.03 and a spillover effect of −0.01.

The school emerges as a crucial factor related to the depopulation of a municipality or vice versa its growth, being naturally both cause and consequence. All other things being equal, the presence of a primary school in the municipality in 2011 is positively linked to an increase in the population in subsequent years. As reported in Table 3 the presence of a primary school in the municipality has a direct effect of increasing the average annual growth rate by 0.76% points, and a spillover effect in the neighbouring municipalities by 0.32% points on average. On the other hand, given the ministerial requirements for the creation of classes, it is the same lack or the reduced number of children residing in the municipality that leads to the closure of primary schools and the movement of families with children to neighbouring municipalities where a school is present. The vicious circle feeds itself, confirming, however, that the possibility of sending children to school is an essential element to stem the depopulation of small municipalities, especially in the internal areas of the country.

Finally, the economic-productive structure in 2011 has a significant relationship with population change; in particular, the share of those employed in agriculture and industry is slightly negatively associated with an increase in population. The direct effect of an increase of 1 point in the percentage of workers in agriculture is to decrease the average annual growth rate of the same municipality in the next ten years by 0.12% points and to decrease the average annual growth rate of the neighbouring municipalities by 0.05% points on average, while for the percentage of workers in the industry the effects are −0.11 and −0.05 on average respectively.

6 Conclusions and Further Developments

It is unquestionable that the European countries are characterized by a very strong territorial heterogeneity, with some regions or areas experiencing population growth and some others a depopulation. According to the latest population projections of regional demographic patterns across the 31 countries (the 27 European Union Member States and the four EFTA countries), two out of three of the NUTS level 3 regions are projected to have a smaller population in 2050 than in 2019. These diverging trends are expected to continue in the future, further worsening the current regional and country demographic inequalities [35].

Our study proved that space matters in defining population growth and decline, underlying the importance of the spatial demography approach in studying such kind of processes [32, 52]. The analysis of the predictors of the average annual growth rate in the last ten years at the municipality level showed a strong effect of the spatial dimension too. The demographic composition of the population is confirmed to have a determinant effect on the dynamics of the next years. Also relevant is the contribution of the socio-economic dimension experienced by individuals whose faster—or at least less slow—transition to adulthood gives a crucial contribution to the growth of the future. The demographic structure of the population (i.e. the percentage of children and of individuals over 75) and the presence of a primary school are revealed to be crucial factors related to the depopulation of the municipality or vice versa, being naturally both cause and consequence. For most of the covariates included in the SAR model, we detect the existence of both direct and indirect effects. This means that the change in the dependent variable (population growth/decline) is associated with both the change of the specific explanatory variable in the municipality (direct effect) and with the change in the same explanatory variable in the neighbouring municipalities. This result confirms the existence of strong spatial diffusion processes at the municipality level.

So far, the recent experience of Covid-19 has shown the limits of distance learning for schools of different levels and particularly for pupils. Starting from our analysis and from the elements that emerged in this health crisis, it is evident that the maintenance of a primary school cannot be neglected if we want to introduce policies to stem the depopulation of the most remote and isolated. On the contrary, for the restart of the social elevator also for those who live in internal areas or remote areas it is crucial to invest in a high quality and full time school that will lay the cultural foundations for new generations. In this perspective, policy makers and local administrators should increase the importance of the territorial capital of each single Italian reality [8, 9] in order to promote a new development both in economic and demographic sense.

The next steps of our research will address two major points. The first one is about the availability of the inter-census reconstructed population (2002–2018) that will allow us to measure the effect of natural and migratory component on the variation of population across time and space. The second one is about the regression model. In the next steps we will try to implement more sophisticated regression models

to evaluate the spatially lagged effect of the covariates, to account for spatial non stationarity of estimations and map local parameters of estimation process [10, 32].

References

1. Akaike, H.: A new look at the statistical model identification. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds.) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY (1974). https://doi.org/10.1007/978-1-4612-1694-0_16
2. Alderotti, G.: Female employment and first childbirth in Italy: what news? Genus **78**(1), 1–19 (2022)
3. Anselin, L.: Local indicators of spatial association—Lisa. Geogr. Anal. **27**(2), 93–115 (1995)
4. Anselin, L.: Spatial econometrics. In: Baltagi, B.H. (ed.) A Companion to Theoretical Econometrics, pp. 310–333. Blackwell Publishing (2001)
5. Anselin, L.: Exploring Spatial Data with GeoDaTM: A Workbook. Center for Spatial Integrated Social Sciences. GeoDa Press, Chicago (2005)
6. Anselin, L., Rey, S.J.: Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and Pysal. GeoDa Press, Chicago (2014)
7. Barbiellini, A., Gomellini, M., Piselli P.: Il contributo della demografia alla crescita economica: duecento anni di "storia" italiana. In: «Questioni di Economia e Finanza» (Occasional Papers), no. 431 (2018). www.bancaditalia.it
8. Benassi, F., D'Elia, M., Petrei, F.: The "meso" dimension of territorial capital: evidence from Italy. Reg. Sci. Policy Pract. **13**(1), 159–175 (2021a)
9. Benassi, F., Busetta, A., Gallo, G., Stranges, M.: Le diseguaglianze tra territori. In: Billari, F.C., Tomassini, C.: (A cura di). AISP – Rapporto sulla popolazione. L'Italia e le sfide della demografia, Il Mulino, Bologna, pp. 135–161 (2021b)
10. Benassi, F., Naccarato, A.: Households in potential economic distress. A geographically weighted regression model for Italy, 2001–2011. Spat. Stat. **21**, 362–376 (2017)
11. Burillo, P., Salvati, L., Matthews, S.A., Benassi, F.: Local-scale fertility variations in a low-fertility country: evidence from Spain (2002–2017). Can. Stud. Popul. **47**(4), 279–295 (2020)
12. Burnham, K.P., Anderson, D.R.: Model Selection and Multi-model Inference: A Practical-Theoretic Approach. Springer, Berlin (2002)
13. Busetta, A., Giambalvo, O.: The effect of women's participation in the labour market on the postponement of first childbirth: a comparison of Italy and Hungary. J. Popul. Res. **31**, 151–192 (2014)
14. Caltabiano, M., Dreassi, E., Rocco, E., Vignoli, D.: A sub-regional analysis of family change: The spatial diffusion of one-parent families across Italian municipalities, 1991–2011. Popul. Space Place **25**, 1–16 (2019)
15. Camagni, R., Gibelli, M.C., Rigamonti, P.: Urban mobility and Urban form: the social and environmental costs of different patterns of urban expansion. Ecol. Econ. **40**(2), 1283–1402 (2002)
16. Cirella, G.T., Russo, A., Benassi, F., Czermański, E., Goncharuk, A.G., Oniszczuk-Jastrzabek, A.: Energy re-shift for an urbanizing world. Energies **14**(17), 5516 (2021). https://doi.org/10.3390/en14175516
17. Coleman, D., Rowthorn, R.: Who's afraid of population decline? A critical examination of its consequences. Popul. Dev. Rev. **37**(s1), 217–248 (2011)
18. De Lucia, M., Benassi, F., Meroni, F., Musacchio, G., Pino, N.A., Strozza, S.: Seismic disasters and the demographic perspectives; 1968, Belice and 1980, Irpinia-Basilicata (Southern Italy) Case Studies. Ann. Geophys. **63**(1) (2020). https://doi.org/10.4401/ag-8298
19. Elshof, H., Haartsen, T., Mulder, C.H.: The effect of primary school absence and closure on inward and outward flows of families. Tijdschrift voor economische en sociale geografie **106**(5), 625–635 (2015)

20. Espon: Shrinking rural regions in Europe Towards smart and innovative approaches to regional development challenges in depopulation rural regions. Policy Brief, Espon Egtc. (2017)
21. European Commission: European Spatial Development Perspective. Towards balanced and sustainable development of the territory of EU. Office for Official Publications of the European Communities, Luxemburg (1999)
22. Glaeser, E.G.: Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier and Happier. Penguin Press, New York (2011)
23. Golgher, A.B., Voss, P.R.: How to interpret the coefficients of spatial models: spillovers, direct and indirect effects. Spat. Demogr. **4**, 175–205 (2016). https://doi.org/10.1007/s40980-015-0016-y
24. Golini, A., Mussino, A., e Savioli, M.: Il malessere demografico in Italia: una ricerca sui comuni italiani, Bologna, Il Mulino (2000)
25. James, P.: Urban Sustainability in Theory and Practice: Circles of Sustainability. Routledge, London (2015)
26. Innocenti, N., Vignoli, D., Lazzeretti, L.: Economic complexity and fertility. Insights from a low fertility country. Reg. Stud. **55**(8), 1388–1402 (2021)
27. Kempen, R.V., Marcuse, P.: A new spatial order in cities? Am. Behav. Sci. **41**(3), 285–298 (1997)
28. Kroismayr, S.: Small school closures in rural areas—the beginning or the end of a downward spiral? Some evidence from Austria. In: Anson, J., Bartl, W., Kulczycki, A. (eds.) Studies in the Sociology of Population, pp. 275–300. Springer, Berlin (2019)
29. Kuczabski, A., Michalski, T.: The process of depopulation in the rural areas of Ukraine. Quaestiones Geographicae **32**(4), 81–90 (2013)
30. Lasanta, T., Arnáez, J., Pascual, N., Ruiz-Flaño, P., Errea, M.P., Lana-Renault, N.: Space–time process and drivers of land abandonment in Europe. Catena (149), 810–823 (2017)
31. LeSage, J., Pace, R.K.: Introduction to Spatial Econometrics. CRC Press, Boca Raton, FL (2009)
32. Matthews, S.A., Parker, D.M.: Progress in spatial demography. Demogr. Res. **28**(10), 271–312 (2013)
33. Mendez, C.: Spatial regression analysis in R. R Studio/Rpubs (2020). https://rpubs.com/quarcs-lab/tutorial-spatial-regression
34. Moran, P.A.P.: The interpretation of statistical maps. J. R. Stat. Soc. **10**(2), 243–251 (1948)
35. Newsham, N., Rowe, F.: Understanding the trajectories of population decline across rural and urban Europe: a sequence analysis, preprint (2022). https://arxiv.org/abs/2203.09798 (03/22/2022)
36. OECD: Divided Cities: Understanding Intra-urban Inequalities. OECD Publishing, Paris (2018). https://doi.org/10.1787/9789264300385-en
37. Potančoková, M., Stonawski, M., Gailey, N.: Migration and demographic disparities in macro-regions of the European Union, a view to 2060. Demogr. Res. **45**, 1317–1354 (2021)
38. Pužulis, A., Kūle, L.: Shrinking of rural territories in Latvia. Economics of the European Union **10** (2016). https://doi.org/10.5755/j01.eis.0.10.14988
39. Rees, P., Van Der Gaag, N., De Beer, J., Heins, F.: European regional populations: current trends, future pathways, and policy options. Eur. J. Popul./Revue Européenne De Démographie **28**(4), 385–416 (2012)
40. Reynaud, C., Miccoli, S. : Depopulation and the aging population: the relationship in Italian municipalities. Sustainability **10**(4) (2018). https://doi.org/10.3390/su10041004
41. Reynaud, C., Miccoli, S., Lagona, F.: Population ageing in Italy: an empirical analysis of change in ageing index across time and space. Spat. Demogr. **6**, 235–251 (2018)
42. Reynaud, C., Miccoli, S., Benassi, F., Naccarato, A., Salvati, L.: Unravelling a demographic 'Mosaic': spatial patterns and contextual factors of depopulation in Italian municipalities, 1981–2011. Ecol. Ind. 115 (2020). https://doi.org/10.1016/j.ecolind.2020.106356
43. Sassen, S.: The global cities: New York, London, Tokyo. Princeton University Press, Princeton, N.J. (1991)

44. Sassen, S.: Cities in the global economy. In: Paddison, R. (ed.) Handbook of Urban Studies, pp. 256–272. Sage, Thousand Oaks, CA (2001)
45. Salvati, L., Benassi, F., Miccoli, S., Rabieri-Dastjerdi, H., Matthews, S.A.: *Spatial* variability of total fertility rate and crude birth rate in a low-fertility country: patterns and trends in regional and local scale heterogeneity across Italy, 2002–2018. Appl. Geogr. **124** (2020). https://doi.org/10.1016/j.apgeog.2020.102321
46. Strozza, S., Benassi, F., Gallo, G., Ferrara, R.: Recent demographic trends in the major Italian urban agglomerations: the role of foreigners. Spat. Demogr. **4**(1), 39–70 (2016)
47. Sun, F., Matthews, S.A., Yang, T.C., Hu, M.H.: A spatial analysis of the COVID-19 period prevalence in US counties through June 28, 2020: where geography matters? Ann. Epidemiol. **52**, 54–59 (2020)
48. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**, 234–240 (1970)
49. Tocchioni, V.: Exploring the childless universe: profiles of women and men without children in Italy. Dem. Res. **38**, 451–470 (2018)
50. Yang, T.C., Noah, A., Shoff, C.: Exploring geographic variation in US mortality rates using a spatial Durbin approach. Popul. Space Place **21**(1), 18–37 (2015)
51. Vanolo, A.: Per uno sviluppo policentrico dello spazio europeo. Sistemi innovativi territoriali nell'Europa sud-occidentale. Milano, Franco Angeli (2003)
52. Voss, P.R.: Demography as a spatial social science. Popul. Res. Policy Rev. **26**, 457–476 (2007)

Model Predictivity Assessment: Incremental Test-Set Selection and Accuracy Evaluation

Elias Fekhari, Bertrand Iooss, Joseph Muré, Luc Pronzato, and Maria-João Rendas

Abstract Unbiased assessment of the predictivity of models learnt by supervised machine learning (ML) methods requires knowledge of the learned function over a reserved test set (not used by the learning algorithm). The quality of the assessment depends, naturally, on the properties of the test set and on the error statistic used to estimate the prediction error. In this work we tackle both issues, proposing a new predictivity criterion that carefully weights the individual observed errors to obtain a global error estimate, and using incremental experimental design methods to "optimally" select the test points on which the criterion is computed. Several incremental constructions are studied, including greedy-packing (coffee-house design), support points and kernel herding techniques. Our results show that the incremental and weighted versions of the latter two, based on Maximum Mean Discrepancy concepts, yield superior performance. An industrial test case provided by the historical French electricity supplier (EDF) illustrates the practical relevance of the methodology, indicating that it is an efficient alternative to expensive cross-validation techniques.

Keywords Design of experiments · Discrepancy · Gaussian process · Machine learning · Metamodel · Validation

E. Fekhari · B. Iooss (✉) · J. Muré
EDF R&D, 6 Quai Watier, 78401 Chatou, France
e-mail: bertrand.iooss@edf.fr

E. Fekhari
e-mail: elias.fekhari@edf.fr

J. Muré
e-mail: joseph.mure@edf.fr

L. Pronzato · M.-J. Rendas
CNRS, Université Côte d'Azur, Laboratoire I3S, Bât. Euclide, Les Algorithmes, 2000 route des Lucioles, 06900 Sophia Antipolis cedex, France
e-mail: luc.pronzato@i3s.unice.fr

M.-J. Rendas
e-mail: rendas@i3s.unice.fr

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_20

1 Introduction

The development of tools for automatic diagnosis relying on learned models imposes strict requirements on model validation. For example, in industrial non-destructive testing (e.g. for the aeronautic or the nuclear industry), generalized automated inspection, which increases efficiency and lowers costs, must provide high performance guarantees [14, 20]. Establishing these guarantees requires availability of a reserved test set, i.e. a data set that has not been used either to train or to select the machine learning (ML) model [3, 21, 56]. Using the prediction residuals on this test set, an independent evaluation of the proposed ML model can be done, enabling the estimation of relevant performance metrics, such as the mean-squared error for regression problems, or the misclassification rate for classification problems.

The same need for independent test sets arises in the area of computer experiments, where computationally expensive simulation codes are often advantageously replaced by ML models, called surrogate models (or metamodels) in this context [15, 44]. Such surrogate models can be used, for instance, to estimate the region of the input space that maps to specific values of the model outputs [7] with a significantly decreased computational load when compared to direct use of the original simulation code. Validation of these surrogate models consists in estimating their predictivity, and can either rely on a suitably selected validation sample, or be done by cross-validation [12, 22, 25]. One of the numerical studies presented in this paper will address an example of this situation of practical industrial interest in the domain of nuclear safety assessment, concerning the simulation of thermal-hydraulic phenomena inside nuclear pressurized water reactors, for which finely validated surrogate models have demonstrated their usefulness [28, 31].

In this paper, we present methods to choose a "good" test set, either within a given dataset or within the input space of the model, as recently motivated in [21, 23]. A first choice concerns the size of the test set. No optimal choice exists, and, when only a finite dataset is available, classical ML handbooks [17, 19] provide different heuristics on how to split it, e.g., 80%/20% between the training and test samples, or 50%/25%/25% between the training, validation (used for model selection) and test samples. We shall not formally address this point here (see [56] for a numerical study of this issue), but in the industrial case-study mentioned above we do study the impact of the ratio between the sizes of the training and test sets on the ability of assessing the quality of the surrogate model. A second issue concerns how the test sample is picked within the input space. The simplest—and most common—way to build a test sample is to extract an independent Monte Carlo sample [19]. For small test sets, these randomly chosen points may fall too close to the training points or leave large areas of the input space unsampled, and a more constructive method to select points inside the input domain is therefore preferable. Similar concerns motivate the use of space-filling designs when choosing a small set of runs for cpu-time expensive computer experiments on which a model will be identified [15, 38].

When the test set must be a subset of an initial dataset, the problem amounts to selecting a certain number of points within a finite collection of points. A review of

classical methods for solving this issue is given in [3]. For example, the CADEX and DUPLEX algorithms [24, 50] can sequentially extract points from a database to include them in a test sample, using an inter-point distance criterion. In ML, identifying within the dataset "prototypes" (set of data instances representative of the whole data set) and "criticisms" (data instances poorly represented by the prototypes) has recently been proposed to help model interpretation [32]; the extraction of prototypes and criticisms relies on a Maximum Mean Discrepancy (MMD) criterion [49] (see e.g. [40], and [37] for a performance analysis of greedy algorithms for MMD minimization).

Several algorithms have also been proposed for the case where points need to be added to an already existing training sample. When the goal is to assess the quality of a model learnt using a known training set, one may be tempted to locate the test points the furthest away from the training samples, such that, in some sense, the union of the training and test sets is space-filling. As this paper shows, test sets built in this manner do enable a good assessment of the quality of models learnt with the training set if the observed residuals are appropriately weighted. Moreover, the incremental augmentation of a design can be useful when the assessed model turns out to be of poor quality, or when an additional computational budget is available after a first study [46, 47]. Different empirical strategies have been proposed for incremental space-filling design [8, 22, 27], which basically entail the addition of new points in the zones poorly covered by the current design. Shang and Apley [46] have recently proposed an improvement of the CADEX algorithm, called the Fully-sequential space-filling (FSSF) design; see also [36] for an alternative version of coffee-house design enforcing boundary avoidance. Although they are developed for different purposes, nested space filling designs [41] and sliced space filling designs [42] can also be used to build sequential designs.

In this work, we provide new insights into these subjects in two main directions: (i) definition of new predictivity criteria through an optimal weighting of the test points residuals, and (ii) use of test sets built by incremental space-filling algorithms, namely FSSF, support points [29] and kernel herding [6], the latter two algorithms being typically used to provide a representative sample of a desired theoretical or empirical distribution. Besides, this paper presents a numerical benchmark analysis comparing the behaviour of the three algorithms on a selected set of test cases.

This paper is organized as follows. Section 2 defines the predictivity criterion considered and proposes different methods for its estimation. Section 3 presents the three algorithms used for test-point selection: FSSF, support points and kernel herding. Our numerical results are presented in Sects. 4 and 5: in Sect. 4 a test set is freely chosen within the entire input space, while in Sect. 5 an existing data set can be split into a training sample and a test set. Finally, Sect. 6 concludes and outlines some perspectives.

2 Predictivity Assessment Criteria for an ML Model

In this section, we propose a new criterion to assess the predictive performance of a model, derived from a standard model quality metric by suitably weighting the errors observed on the test set. We denote by $\mathcal{X} \subset \mathbb{R}^d$ the space of the input variables $\mathbf{x} = (x_1, \ldots, x_d)$ of the model. Then let $y(\mathbf{x}) \in \mathbb{R}$ (resp. $y(\mathbf{x}') \in \mathbb{R}$) be the observed output at point $\mathbf{x} \in \mathcal{X}$ (resp. $\mathbf{x}' \in \mathcal{X}$). We denote by $(\mathbf{X}_m, \mathbf{y}_m)$ the training sample, with $\mathbf{y}_m = [y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(m)})]^\top$. The test sample is denoted by $(\mathbf{X}_n, \mathbf{y}_n) = (\mathbf{x}^{(m+i)}, y(\mathbf{x}^{(m+i)}))_{1 \le i \le n}$.

2.1 The Predictivity Coefficient

Let $\eta_m(\mathbf{x})$ denote the prediction at point $\mathbf{x}$ of a model learned using $(\mathbf{X}_m, \mathbf{y}_m)$ [19, 43]. A classical measure for assessing the predictive ability of η_m, in order to evaluate its validity, is the predictivity coefficient. Let μ denote the measure that weights how comparatively important it is to accurately predict y over the different regions of $\mathcal{X}$. For example the input could be a random vector with known distribution: in that case, this distribution would be a reasonable choice for μ. The true (ideal) value of the predictivity is defined as the following normalization of the Integrated Square Error (ISE):

$$Q^2_{\text{ideal}}(\mu) = 1 - \frac{\text{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)}{V_\mu}\,, \tag{1}$$

where

$$\begin{aligned} \text{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m) &= \int_{\mathcal{X}} [y(\mathbf{x}) - \eta_m(\mathbf{x})]^2 \, \mathrm{d}\mu(\mathbf{x})\,, \\ V_\mu &= \int_{\mathcal{X}} \left[y(\mathbf{x}) - \int_{\mathcal{X}} y(\mathbf{x}') d\mu(\mathbf{x}') \right]^2 \mathrm{d}\mu(\mathbf{x})\,. \end{aligned}$$

The ideal predictivity $Q^2_{\text{ideal}}(\mu)$ is usually estimated by its empirical version calculated over the test sample $(\mathbf{X}_n, \mathbf{y}_n)$, see [10, p. 32]:

$$\widehat{Q}^2_n = 1 - \frac{\sum_{\mathbf{x} \in \mathbf{X}_n} [y(\mathbf{x}) - \eta_m(\mathbf{x})]^2}{\sum_{\mathbf{x} \in \mathbf{X}_n} \left[y(\mathbf{x}) - \overline{y}_n \right]^2}\,, \tag{2}$$

where $\overline{y}_n = (1/n) \sum_{i=1}^n y(\mathbf{x}^{(m+i)})$ denotes the empirical mean of the observations in the test sample. Note that the calculation of $\widehat{Q}^2_n$ only requires access to the predictor $\eta_m(\cdot)$. To compute $\widehat{Q}^2_n$, one does not need to know the training set which was used to build $\eta_m(\cdot)$. $\widehat{Q}^2_n$ is the coefficient of determination (a standard notion in parametric regression) common in prediction studies [22, 25], often called "Nash-Sutcliffe cri-

terion" [35]: it compares the prediction errors obtained with the model η_m with those obtained when prediction equals the empirical mean of the observations. Thus, the closer $\widehat{Q}_n^2$ is to one, the more accurate the surrogate model is (for the test set considered). On the contrary, $\widehat{Q}_n^2$ close to zero (negative values are possible too) indicates poor predictions abilities, as there is little improvement compared to prediction by the simple empirical mean of the observations. The next section shows how a suitable weighting of the residual on the training sample may be key to improving the estimation of $\widehat{Q}_n^2$.

2.2 Weighting the Test Sample

Let $\xi_n = n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}^{(i)}}$ be the empirical distribution of the prediction error, with $\delta_{\mathbf{x}}$ the Dirac measure at $\mathbf{x}$. In $\widehat{Q}_n^2$, $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$ is estimated by the empirical average of the squared residuals

$$\mathrm{ISE}_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m) = \frac{1}{n} \sum_{i=1}^{n} \left[y(\mathbf{x}^{(m+i)}) - \eta_m(\mathbf{x}^{(m+i)}) \right]^2 .$$

When the points $\mathbf{x}^{(m+i)}$ of the test set $\mathbf{X}_n$ are distant from the points of the training set $\mathbf{X}_m$, the squared prediction errors $|y(\mathbf{x}^{(m+i)}) - \eta_m(\mathbf{x}^{(m+i)})|^2$ tend to represent the worst possible error situations, and $\mathrm{ISE}_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m)$ tends to overestimate $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$. In this section, we postulate a statistical model for the prediction errors in order to be able to quantify this potential bias when sampling the residual process, enabling its subsequent correction .

In [39], the authors propose a weighting scheme for the test set when the ML model interpolates the train set observations. They suggest several variants corresponding to different constraints on the weights (e.g., non-negativity, summing to one). In the following, we consider the unconstrained version only, which in our experience works best. Let $\delta_m(\mathbf{x}) = y(\mathbf{x}) - \eta_m(\mathbf{x})$ denote the predictor error and assume it is a realization of a Gaussian Process (GP) with zero mean and covariance kernel $\sigma^2 K_{|m}$, which we shall note $\delta_m(\mathbf{x}) \sim \mathsf{GP}(0, \sigma^2 K_{|m})$, with

$$\sigma^2 K_{|m}(\mathbf{x}, \mathbf{x}') = \mathbb{E}\{\delta_m(\mathbf{x})\delta_m(\mathbf{x}')\} = \sigma^2 \left[K(\mathbf{x}, \mathbf{x}') - \mathbf{k}_m^\top(\mathbf{x}) \mathbf{K}_m^{-1} \mathbf{k}_m(\mathbf{x}') \right] .$$

Here, $\mathbf{k}_m(\mathbf{x})$ denotes the column vector $[K(\mathbf{x}, \mathbf{x}^{(1)}), \ldots, K(\mathbf{x}, \mathbf{x}^{(m)})]^\top$ and $\mathbf{K}_m$ is the $m \times m$ matrix whose element (i, j) is given by $\{\mathbf{K}_m\}_{i,j} = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, with K a positive definite kernel. The rationale for using this model is simple: assume first a prior GP model $\mathsf{GP}(0, \sigma^2 K)$ for the error process $\delta(\mathbf{x})$; if η_m interpolates the observations $\mathbf{y}_m$, the errors observed at the design points $\mathbf{x}^{(i)}$ equal zero, $i = 1, \ldots, m$, leading finally to the posterior $\mathsf{GP}(0, \sigma^2 K_{|m})$ for $\delta_m(\mathbf{x})$.

However, the predictor η_m is not always an interpolator, see Sect. 5 for an example, so we extend the approach of [39] to the general situation where η_m does not

necessarily interpolate the training data $\mathbf{y}_m$. The same prior $\mathsf{GP}(0, \sigma^2 K)$ for $\delta(\mathbf{x})$ yields $\delta_m(\mathbf{x}) \sim \mathsf{GP}(\widehat{\delta}_m(\mathbf{x}), \sigma^2 K_{|m})$, where

$$\widehat{\delta}_m(\mathbf{x}) = \mathbf{k}_m^\top(\mathbf{x}) \mathbf{K}_m^{-1}(\mathbf{y}_m - \eta_m) \tag{3}$$

is the Kriging interpolator for the errors, with $\eta_m = [\eta_m(\mathbf{x}^{(1)}), \ldots, \eta_m(\mathbf{x}^{(m)})]^\top$.

The model above allows us to study how well $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$ is estimated using a given test set. Denote by $\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m)$ the expected squared error when estimating $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$ by $\mathrm{ISE}_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m)$,

$$\begin{aligned}
\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m) &= \mathbb{E}\left\{\left[\mathrm{ISE}_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m) - \mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)\right]^2\right\} \\
&= \mathbb{E}\left\{\left[\int_{\mathcal{X}} \delta_m^2(\mathbf{x})\, \mathrm{d}(\xi_n - \mu)(\mathbf{x})\right]^2\right\} \\
&= \mathbb{E}\left\{\int_{\mathcal{X}^2} \delta_m^2(\mathbf{x}) \delta_m^2(\mathbf{x}')\, \mathrm{d}(\xi_n - \mu)(\mathbf{x}) \mathrm{d}(\xi_n - \mu)(\mathbf{x}')\right\}.
\end{aligned}$$

Tonelli's theorem gives

$$\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m) = \int_{\mathcal{X}^2} \mathbb{E}\{\delta_m^2(\mathbf{x}) \delta_m^2(\mathbf{x}')\}\, \mathrm{d}(\xi_n - \mu)(\mathbf{x}) \mathrm{d}(\xi_n - \mu)(\mathbf{x}').$$

Since $\mathbb{E}\left\{U^2 V^2\right\} = 2\ (\mathbb{E}\{UV\})^2 + \mathbb{E}\left\{U^2\right\} \mathbb{E}\left\{V^2\right\}$ for any one-dimensional normal centered random variables U and V, when $\eta_m(\mathbf{x})$ interpolates $\mathbf{y}_m$, we obtain

$$\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m) = \sigma^2\, d^2_{\overline{K}_{|m}}(\xi_n, \mu), \tag{4}$$

where

$$\begin{aligned}
d^2_{\overline{K}_{|m}}(\xi_n, \mu) &= \int_{\mathcal{X}^2} \overline{K}_{|m}(\mathbf{x}, \mathbf{x}')\, \mathrm{d}(\xi_n - \mu)(\mathbf{x}) \mathrm{d}(\xi_n - \mu)(\mathbf{x}'), \\
\overline{K}_{|m}(\mathbf{x}, \mathbf{x}') &= 2\, K^2_{|m}(\mathbf{x}, \mathbf{x}') + K_{|m}(\mathbf{x}, \mathbf{x}) K_{|m}(\mathbf{x}', \mathbf{x}'),
\end{aligned} \tag{5}$$

and we recognize $d^2_{\overline{K}_{|m}}(\xi_n, \mu)$ as the squared Maximum-Mean-Discrepancy (MMD) between ξ_n and μ for the kernel $\overline{K}_{|m}$; see (18) in Appendix A. Note that σ^2 only appears as a multiplying factor in (4), with the consequence that σ^2 does not impact the choice of a suitable ξ_n.

When η_m does not interpolate $\mathbf{y}_m$, similar developments still give $\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m) = \sigma^2\, d^2_{\overline{K}_{|m}}(\xi_n, \mu)$, with now

$$\begin{aligned}
\overline{K}_{|m}(\mathbf{x}, \mathbf{x}') = 2 &\left[K_{|m}(\mathbf{x}, \mathbf{x}') + 2\, \widehat{\delta}_m(\mathbf{x}) \widehat{\delta}_m(\mathbf{x}')\right] K_{|m}(\mathbf{x}, \mathbf{x}') \\
&+ \left[\widehat{\delta}_m^2(\mathbf{x}) + K_{|m}(\mathbf{x}, \mathbf{x})\right] \left[\widehat{\delta}_m^2(\mathbf{x}') + K_{|m}(\mathbf{x}', \mathbf{x}')\right],
\end{aligned}$$

where $\widehat{\delta}_m(\mathbf{x})$ is given by (3).

The idea is to replace ξ_n, uniform on $\mathbf{X}_n$, by a nonuniform measure ζ_n supported on $\mathbf{X}_n$, $\zeta_n = \sum_{i=1}^n w_i \delta_{\mathbf{x}^{(m+i)}}$ with weights $\mathbf{w}_n = (w_1, \ldots, w_n)^\top$ chosen such that the estimation error $\overline{\Delta}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m)$, and thus $d^2_{\overline{K}_{|m}}(\zeta_n, \mu)$, is minimized. Direct calculation gives

$$d^2_{\overline{K}_{|m}}(\zeta_n, \mu) = \mathscr{E}_{\overline{K}_{|m}}(\mu) - 2\,\mathbf{w}_n^\top \mathbf{p}_{\overline{K}_{|m},\mu}(\mathbf{X}_n) + \mathbf{w}_n^\top \overline{\mathbf{K}}_{|m}(\mathbf{X}_n)\mathbf{w}_n\,,$$

where $\mathbf{p}_{\overline{K}_{|m},\mu}(\mathbf{X}_n) = \left[P_{\overline{K}_{|m},\mu}(\mathbf{x}^{(m+1)}), \ldots, P_{\overline{K}_{|m},\mu}(\mathbf{x}^{(m+n)})\right]^\top$, with $P_{\overline{K}_{|m},\mu}(\mathbf{x})$ defined by (17) in Appendix A, $\{\overline{\mathbf{K}}_{|m}(\mathbf{X}_n)\}_{i,j} = \overline{K}_{|m}(\mathbf{x}^{(m+i)}, \mathbf{x}^{(m+j)})$ for any $i, j = 1, \ldots, n$, and $\mathscr{E}_{\overline{K}_{|m}}(\mu) = \int_{\mathcal{X}^2} \overline{K}_{|m}(\mathbf{x}, \mathbf{x}')\,\mathrm{d}\mu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}')$.

When $\mathbf{X}_m \cap \mathbf{X}_n = \emptyset$, the $n \times n$ matrix $\mathbf{K}_{|m}(\mathbf{X}_n)$, whose element i, j equals $K_{|m}(\mathbf{x}^{(m+i)}, \mathbf{x}^{(m+j)})$, is positive definite. The elementwise (Hadamard) product $\mathbf{K}_{|m}(\mathbf{X}_n) \circ \mathbf{K}_{|m}(\mathbf{X}_n)$ is thus positive definite too, implying that $\overline{\mathbf{K}}_{|m}(\mathbf{X}_n)$ is positive definite. The optimal weights $\mathbf{w}_n^*$ minimizing $d^2_{\overline{K}_{|m}}(\zeta_n, \mu)$ are thus

$$\mathbf{w}_n^* = \overline{\mathbf{K}}^{-1}_{|m}(\mathbf{X}_n)\mathbf{p}_{\overline{K}_{|m},\mu}(\mathbf{X}_n)\,. \tag{6}$$

We shall denote by ζ_n^* the measure supported on $\mathbf{X}_n$ with the optimal weights (6) and

$$\begin{aligned} Q^2_{n*} &= 1 - \frac{\mathrm{ISE}_{\zeta_n^*}(\mathbf{X}_m, \mathbf{y}_m)}{\frac{1}{n}\sum_{i=1}^n \left[y(\mathbf{x}^{(m+i)}) - \overline{y}_n\right]^2} \\ &= 1 - \frac{\sum_{i=1}^n w_i^* \left[y(\mathbf{x}^{(m+i)}) - \eta_m(\mathbf{x}^{(m+i)})\right]^2}{\frac{1}{n}\sum_{i=1}^n \left[y(\mathbf{x}^{(m+i)}) - \overline{y}_n\right]^2}\,, \end{aligned} \tag{7}$$

with $\overline{y}_n = (1/n)\sum_{i=1}^n y(\mathbf{x}^{(m+i)})$. Notice that the weights w_i^* do not depend on the variance parameter σ^2 of the GP model.

Remark 1 When $\mathbf{X}_n$ is constructed by kernel herding, see Sect. 3.3, K can be chosen identical to the kernel used there. This will be the case in Sects. 4 and 5, but it is not mandatory.

Conversely, one may think of choosing a design $\mathbf{X}_n$ that minimizes $d^2_{\overline{K}_{|m}}(\xi_n, \mu)$, or $d^2_{\overline{K}_{|m}}(\zeta_n^*, \mu)$, also with the objective to obtain a precise estimation of $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$. This can be achieved by kernel herding using the kernel $\overline{K}_{|m}$ and is addressed in [39]. However, the numerical results presented there show that the precise choice of the test set $\mathbf{X}_n$ has a marginal effect compared to the effect of non-uniform weighting with $\mathbf{w}_n^*$, provided that $\mathbf{X}_n$ fills the holes left in $\mathcal{X}$ by the training design $\mathbf{X}_m$. ◁

Remark 2 When the observations $y(\mathbf{x}^{(i)})$, $i = 1, \ldots, n$, are available at the validation stage, an alternative version of $\widehat{Q}^2_n$ would be

$$\widehat{Q}_n'^2 = 1 - \frac{\sum_{\mathbf{x}\in\mathbf{X}_n} [y(\mathbf{x}) - \eta_m(\mathbf{x})]^2}{\sum_{\mathbf{x}\in\mathbf{X}_n} \left[y(\mathbf{x}) - \overline{y}_m\right]^2}, \tag{8}$$

where $\overline{y}_m = (1/m) \sum_{i=1}^m y(\mathbf{x}^i)$, which compares the performance on the test set of two predictors η_m and $\overline{y}_m$ based on the same training set. It is then possible to also apply a weighting procedure to the *denominator* of $\widehat{Q}_n'^2$,

$$D_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m) = \frac{1}{n} \sum_{i=1}^n \left[y(\mathbf{x}^{(m+i)}) - \overline{y}_m\right]^2 ,$$

in order to make it resemble its idealized version $V'_\mu(\mathbf{y}_m) = \int_{\mathcal{X}} \left[y(\mathbf{x}) - \overline{y}_m\right]^2 \mathrm{d}\mu(\mathbf{x})$. The GP model is now $\varepsilon_m(\mathbf{x}) = y(\mathbf{x}) - \overline{y}_m \sim \mathsf{GP}(\eta_m(\mathbf{x}) - \overline{y}_m, \sigma^2 K_{|m})$. Similar developments to those used above for the numerator of $\widehat{Q}_n^2$ yield

$$\begin{aligned}
\overline{\Delta'}^2(\xi_n, \mu; \mathbf{X}_m, \mathbf{y}_m) &= \mathbb{E}\left\{\left[D_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m) - V_\mu(\mathbf{y}_m)\right]^2\right\} \\
&= \mathbb{E}\left\{\int_{\mathcal{X}^2} \varepsilon_m^2(\mathbf{x})\varepsilon_m^2(\mathbf{x}')\, \mathrm{d}(\xi_n - \mu)(\mathbf{x})\mathrm{d}(\xi_n - \mu)(\mathbf{x}')\right\} \\
&= \sigma^2\, d^2_{\overline{K'}_{|m}}(\xi_n, \mu)\,,
\end{aligned}$$

where

$$\begin{aligned}
\overline{K'}_{|m}(\mathbf{x}, \mathbf{x}') = \overline{K}_{|m}(\mathbf{x}, \mathbf{x}') &+ [\eta_m(\mathbf{x}) - \overline{\mathbf{y}}_m]^2[\eta_m(\mathbf{x}') - \overline{\mathbf{y}}_m]^2 \\
&+ [\eta_m(\mathbf{x}) - \overline{\mathbf{y}}_m]^2 K_{|m}(\mathbf{x}', \mathbf{x}') + [\eta_m(\mathbf{x}') - \overline{\mathbf{y}}_m]^2 K_{|m}(\mathbf{x}, \mathbf{x}) \\
&+ 4\,[\eta_m(\mathbf{x}) - \overline{\mathbf{y}}_m][\eta_m(\mathbf{x}') - \overline{\mathbf{y}}_m] K_{|m}(\mathbf{x}, \mathbf{x}')\,.
\end{aligned}$$

We can then substitute $D_{\zeta_n'^*}(\mathbf{X}_m, \mathbf{y}_m)$ for $D_{\xi_n}(\mathbf{X}_m, \mathbf{y}_m)$ in (8), where $\zeta_n'^*$ allocates the weights $\mathbf{w}_n'^* = \overline{\mathbf{K}'}_{|m}^{-1}(\mathbf{X}_n)\mathbf{p}_{\overline{K'}_{|m},\mu}(\mathbf{X}_n)$ to the n points in $\mathbf{X}_n$, with $\{\overline{\mathbf{K}'}_{|m}(\mathbf{X}_n)\}_{i,j} = \overline{K'}_{|m}(\mathbf{x}^{(m+i)}, \mathbf{x}^{(m+j)})$, $i, j = 1, \ldots, n$. ◁

3 Test-set Construction

In the previous section we assumed the test set as given, and proposed a method to estimate $\mathrm{ISE}_\mu(\mathbf{X}_m, \mathbf{y}_m)$ by a weighted sum of the residuals. In this section we address the choice of the test set.

Below we give an overview of the three methods used in this paper, all relying on the concept of space-filling design [15, 38]. While most methods for the construction of such designs choose all points simultaneously, the methods we consider are incremental, selecting one point at a time.

Our objective is to construct an ordered test set of size n, denoted by $\mathbf{X}_n = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\} \subset \mathcal{X}$. When there is no restriction on the choice of $\mathbf{X}_n$, the advantage of using an incremental construction is that it can be stopped once the estimation of the predictivity of an initial model, built with some given design $\mathbf{X}_m$, is considered sufficiently accurate. In case the conclusion is that model predictions are not reliable enough, the full design $\mathbf{X}_{m+n} = \mathbf{X}_m \cup \mathbf{X}_n$ and the associated observations $\mathbf{y}_{m+n}$ can be used to update the model. This updated model can then be tested at additional design points, elements of a new test set to be constructed. All methods presented in this section (except the Fully Sequential Space-Filling method) are implemented in the Python package `otkerneldesign`[1] which is based on the OpenTURNS library for uncertainty quantification [1].

3.1 *Fully-Sequential Space-Filling Design*

The Fully-Sequential Space-Filling forward-reflected (FSSF-fr) algorithm [46] relies on the CADEX algorithm [24] (also called the "coffee-house" method [34]). It constructs a sequence of nested designs in a bounded set $\mathcal{X}$ by sequentially selecting a new point $\mathbf{x}$ as far away as possible from the $\mathbf{x}^{(i)}$ previously selected. New inserted points are selected within a set of candidates $\mathcal{S}$ which may coincide with $\mathcal{X}$ or be a finite subset of $\mathcal{X}$ (which simplifies the implementation, only this case is considered here). The improvement of FSSF-fr when compared to CADEX is that new points are selected *at the same time* far from the previous design points as well as far from the boundary of $\mathcal{X}$.

The algorithm is as follows:

1. Choose $\mathcal{S}$, a finite set of candidate points in $\mathcal{X}$, with size $N \gg n$ in order to allow a fairly dense covering of $\mathcal{X}$. When $\mathcal{X} = [0, 1]^d$, [46] recommends to take $\mathcal{S}$ equal to the first $N = 1\,000\,d + 2\,n$ points of a Sobol sequence in $\mathcal{X}$.
2. Choose the first point $\mathbf{x}^{(1)}$ randomly in $\mathcal{S}$ and define $\mathbf{X}_1 = \{\mathbf{x}^{(1)}\}$.
3. At iteration i, with $\mathbf{X}_i = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}\}$, select

$$\mathbf{x}^{(i+1)} \in \underset{\mathbf{x} \in \mathcal{S} \setminus \mathbf{X}_i}{\operatorname{Arg\,max}} \left[\min \left(\min_{j \in \{1,\ldots,i\}} \|\mathbf{x} - \mathbf{x}^{(j)}\|, \sqrt{2}\, d\, \operatorname{dist}(\mathbf{x}, R(\mathbf{x})) \right) \right], \tag{9}$$

where $R(\mathbf{x})$ is the symmetric of $\mathbf{x}$ with respect to its nearest boundary of $\mathcal{X}$, and set $\mathbf{X}_{i+1} = \mathbf{X}_i \bigcup \mathbf{x}^{(i+1)}$.
4. Stop the algorithm when $\mathbf{X}_n$ has the required size.

The standard coffee-house (greedy packing) algorithm simply uses $\mathbf{x}^{(i+1)} \in \underset{\mathbf{x} \in \mathcal{S} \setminus \mathbf{X}_i}{\operatorname{Arg\,max}} \min_{j \in \{1,\ldots,i\}} \|\mathbf{x} - \mathbf{x}^{(j)}\|$. The role of the reflected point $R(\mathbf{x})$ is to avoid selecting $\mathbf{x}^{(i+1)}$ too close to the boundary of $\mathcal{X}$, which is a major problem with standard coffee-

[1] https://pypi.org/project/otkerneldesign/.

house, especially when $\mathcal{X} = [0, 1]^d$ with d large. The factor $\sqrt{2}d$ in (9) proposed in [46] sets a balance between distance to the design $\mathbf{X}_i$ and distance to the boundary of $\mathcal{X}$. Another scaling factor, depending on the target design size n is proposed in [36].

FSSF-fr is entirely based on geometric considerations and implicitly assumes that the selected set of points should cover $\mathcal{X}$ evenly. However, in the context of uncertainty quantification [48] it frequently happens that the distribution μ of the model inputs is not uniform. It is then desirable to select a test set representative of μ. This can be achieved through the inverse probability integral transform: FSSF-fr constructs $\mathbf{X}_n$ in the unit hypercube $[0, 1]^d$, and an "isoprobabilistic" transform $T : [0, 1]^d \to \mathcal{X}$ is then applied to the points in $\mathbf{X}_i$, T being such that, if $\boldsymbol{U}$ is a random variable uniform on $[0, 1]^d$, then $T(\boldsymbol{U})$ follows the target distribution μ. The transformation can be applied to each input separately when μ is the product of its marginals, a situation considered in our second test-case of Sect. 4, but is more complicated in other cases, see [26, Chap. 4]. Note that FSSF-fr operates in the bounded set $[0, 1]^d$ even if the support of μ is unbounded. The other two algorithms presented in this section are able to directly choose points representative of a given distribution μ and do not need to resort to such a transformation.

3.2 Support Points

Support points [29] are such that their associated empirical distribution ξ_n has minimum Maximum-Mean-Discrepancy (MMD) with respect to μ for the energy-distance kernel of Székely and Rizzo [52, 53],

$$K_E(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\left(\|\mathbf{x}\| + \|\mathbf{x}'\| - \|\mathbf{x} - \mathbf{x}'\|\right). \tag{10}$$

The squared MMD between ξ_n and μ for the distance kernel equals

$$d^2_{K_E}(\xi_n, \mu) = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\|\mathbf{x}^{(i)} - \zeta\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| - \mathbb{E}\|\zeta - \zeta'\|, \tag{11}$$

where ζ and ζ' are independently distributed with μ; see [45]. A key property of the energy-distance kernel is that it is characteristic [51]: for any two probability distributions μ and ξ on $\mathcal{X}$, $d^2_{K_E}(\mu, \xi)$ equals zero if and only if $\mu = \xi$, and so it defines a norm in the space of probability distributions. Compared to more heuristic methods for solving quantization problems, support points benefit from the theoretical guarantees of MMD minimization in terms of convergence of ξ_n to μ as $n \to \infty$.

As $\mathbb{E}\|\mathbf{x}^{(i)} - \zeta\|$ is not known explicitly, in practice μ is replaced by its empirical version μ_N for a given large-size sample $(\mathbf{x}'^{(k)})_{k=1\ldots N}$. The support points $\mathbf{X}^s_n$ are then given by

$$\mathbf{X}_n^s \in \underset{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}}{\text{Arg min}} \left(\frac{2}{nN} \sum_{i=1}^{n} \sum_{k=1}^{N} \|\mathbf{x}^{(i)} - \mathbf{x}'^{(k)}\| - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \right) . \quad (12)$$

The function to be minimized can be written as a difference of functions convex in $\mathbf{X}_n$, which yields a difference-of-convex program. In [29], a majorization-minimization procedure, efficiently combined with resampling, is applied to the construction of large designs (up to $n = 10^4$) in high dimensional spaces (up to $d = 500$). The examples treated clearly show that support points are distributed in a way that matches μ more closely than Monte-Carlo and quasi-Monte Carlo samples [15].

The method can be used to split a dataset into a training set and a test set [23]: the N points $\mathbf{X}_N$ in (12) are those from the dataset, $\mathbf{X}_n^s$ gives the test set and the other $N - n$ points are used for training. There is a serious additional difficulty though, as choosing $\mathbf{X}_n^s$ among the dataset corresponds to a difficult combinatorial optimization problem. A possible solution is to perform the optimization in a continuous domain $\mathcal{X}$ and then choose $\mathbf{X}_n^s$ that corresponds to the closest points in $\mathbf{X}_N$ (for the Euclidean distance) to the continuous solution obtained [23].

The direct determination of support points through (12) does not allow the construction of a nested sequence of test sets. One possibility would be to solve (12) sequentially, one point at a time, in a continuous domain, and then select the closest point within $\mathbf{X}_N$ as the one to be included in the test set. We shall use a different approach here, based on the greedy minimization of the MMD (11) for the candidate set $\mathcal{S} = \mathbf{X}_N$: at iteration i, the algorithm chooses

$$\mathbf{x}_{i+1}^s \in \underset{\mathbf{x}\in\mathcal{S}}{\text{Arg min}} \left(\frac{1}{N} \sum_{k=1}^{N} \|\mathbf{x} - \mathbf{x}'^{(k)}\| - \frac{1}{i+1} \sum_{j=1}^{i} \|\mathbf{x} - \mathbf{x}^{(j)}\| \right) . \quad (13)$$

The method requires the computation of the $N(N-1)/2$ distances $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|$, $i, j = 1, \ldots, N$, $i \neq j$, which hinders its applicability to large-scale problems (a test-case with $N = 1\,000$ is presented in Sect. 5). Note that we consider support points in the input space $\mathcal{X}$ only, with $\mathcal{X} \subseteq \mathbb{R}^d$, in contrast with [23] which considers couples $(\mathbf{x}^{(i)}, y(\mathbf{x}^{(i)}))$ in $\mathbb{R}^{d+1}$ to split a given dataset into a training set and a test set.

Greedy MMD minimization can be applied to other kernels than the distance kernel (10), see [37, 54]. In the next section we consider the closely related method of Kernel Herding (KH) [6], which corresponds to a conditional-gradient descent in the space of probability measures supported on a candidate set $\mathcal{S}$; see, e.g., [40] and the references therein.

3.3 Kernel Herding

Let K be a positive definite kernel on $\mathcal{X} \times \mathcal{X}$. At iteration i of kernel herding, with $\xi_i = (1/i) \sum_{j=1}^{i} \delta_{\mathbf{x}^{(i)}}$ the empirical measure for $\mathbf{X}_i$, the next point $\mathbf{x}_{i+1}$ minimizes

the directional derivative $F_K(\xi_i, \mu, \delta_{\mathbf{x}})$ of the squared MMD $d_K^2(\xi, \mu)$ at $\xi = \xi_i$ in the direction of the delta measure $\delta_{\mathbf{x}}$, see Appendix A. Direct calculation gives $F_K(\xi_i, \mu, \delta_{\mathbf{x}}) = 2\,[P_{K,\xi}(\mathbf{x}) - P_{K,\mu}(\mathbf{x})] - 2\int_{\mathcal{X}}[P_{K,\xi}(\mathbf{x}) - P_{K,\mu}(\mathbf{x})]\,\mathrm{d}\xi(\mathbf{x})$, with $P_{K,\xi}(\mathbf{x})$ (resp. $P_{K,\mu}(\mathbf{x})$) the potential of ξ (resp. μ) at $\mathbf{x}$, see (17), and thus

$$\mathbf{x}_{i+1} \in \operatorname*{Arg\,min}_{\mathbf{x}\in\mathcal{S}} \left[P_{K,\xi_i}(\mathbf{x}) - P_{K,\mu}(\mathbf{x})\right], \tag{14}$$

with $\mathcal{S} \subseteq \mathcal{X}$ a given candidate set. Here, $P_{K,\xi_i}(\mathbf{x}) = (1/i)\sum_{j=1}^{i} K(\mathbf{x}, \mathbf{x}^{(j)})$. When an empirical measure μ_N based on a sample $(\mathbf{x}'^{(k)})_{k=1\ldots N}$ is substituted for μ, we get $P_{K,\mu_N}(\mathbf{x}) = (1/N)\sum_{k=1}^{N} K(\mathbf{x}, \mathbf{x}'^{(k)})$, which gives

$$\mathbf{x}_{i+1} \in \operatorname*{Arg\,min}_{\mathbf{x}\in\mathcal{S}} \left[\frac{1}{i}\sum_{j=1}^{i} K(\mathbf{x}, \mathbf{x}^{(j)}) - \frac{1}{N}\sum_{k=1}^{N} K(\mathbf{x}, \mathbf{x}'^{(k)})\right].$$

When K is the energy-distance kernel (10) we thus obtain (13) with a factor $1/i$ instead of $1/(i+1)$ in the second sum.

The candidate set $\mathcal{S}$ in (14) is arbitrary and can be chosen as in Sect. 3.1. A neat advantage of kernel herding over support points in that the potential $P_{K,\mu}(\mathbf{x})$ is sometimes explicitly available. When $\mathcal{S} = \mathbf{X}_N$, this avoids the need to calculate the $N(N-1)/2$ distances $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|$ and thus allows application to very large sample sizes. This is the case in particular when $\mathcal{X}$ is the cross product of one-dimensional sets $\mathcal{X}_{[i]}$, $\mathcal{X} = \mathcal{X}_{[1]} \times \cdots \times \mathcal{X}_{[d]}$, μ is the product of its marginals $\mu_{[i]}$ on the $\mathcal{X}_{[i]}$, K is the product of one-dimensional kernels $K_{[i]}$, and the one-dimensional integral in $P_{K_{[i]},\mu_{[i]}}(x)$ is known explicitly for each $i \in \{1, \ldots, d\}$. Indeed, for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}$, we then have $P_{K,\mu}(\mathbf{x}) = \prod_{i=1}^{d} P_{K_{[i]},\mu_{[i]}}(x_i)$; see [40]. When K is the product of Matérn kernels with regularity parameter $5/2$ and correlation lengths θ_i, $K(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} K_{5/2,\theta_i}(x_i - x_i')$, with

$$K_{5/2,\theta}(x - x') = \left(1 + \frac{\sqrt{5}}{\theta}|x - x'| + \frac{5}{3\theta^2}(x - x')^2\right)\exp\left(-\frac{\sqrt{5}}{\theta}|x - x'|\right), \tag{15}$$

the one-dimensional potentials are given in Appendix B for $\mu_{[i]}$ uniform on $[0, 1]$ or $\mu_{[i]}$ the standard normal $\mathcal{N}(0, 1)$. When no observation is available, which is the common situation at the design stage, the correlation lengths have to be set to heuristic values. We empirically found the values of the correlation lengths to have a large influence over the design. A reasonable choice for $\mathcal{X} = [0, 1]^d$ is $\theta_i = n^{-1/d}$ for all i, with n the target number of design points; see [40].

3.4 Numerical Illustration

We apply FSSF-fr (denoted FSSF in the following), support points and kernel herding algorithms to the situation where a given initial design of size m has to be completed by a series of additional points $\mathbf{x}^{(m+1)}, \ldots, \mathbf{x}^{(m+n)}$. The objective is to obtain a full design $\mathbf{X}_{m+n}$ that is a good quantization of a given distribution μ.

Figures 1 and 2 correspond to μ uniform on $[0, 1]^2$ and μ the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, with $\mathbf{I}_2$ the 2-dimensional identity matrix, respectively. All methods are applied to the same candidate set $\mathcal{S}$.

The initial designs $\mathbf{X}_m$ are chosen in the class of space-filling designs, well suited to initialize sequential learning strategies [44]. When μ is uniform, the initial design is a maximin Latin hypercube design [33] with $m = 10$ and the candidate set is given by the $N = 2^{12}$ first points $\mathbf{S}_N$ of a Sobol sequence in $[0, 1]$. When μ is normal, the inverse probability transform method is first applied to $\mathbf{S}_N$ and $\mathbf{X}_m$ (this does not raise any difficulty here as μ is the product of its marginals). The candidate points $\mathcal{S}$ are marked in gray on Figs. 1 and 2 and the initial design is indicated by the red crosses. The index i of each added test point $\mathbf{x}^{(m+i)}$ is indicated (the font size decreases with i). In such a small dimension ($d = 2$), a visual appreciation gives the impression that the three methods have comparable performance. We can notice, however, that FSSF tends to choose points closer to the boundary of $\mathcal{S}$ than the other two, and that support points seem to sample more freely the holes of $\mathbf{X}_m$ than kernel herding, which seems to be closer to a space-filling continuation of the training set. We will come back to these designs when analysing the quality of the resulting predictivity metric estimators in the next section.

4 Numerical Results I: Construction of a Training Set and a Test Set

This section presents numerical results obtained on three different test-cases, in dimension 2 (test-cases 1 and 2) and 8 (test-case 3), for which $y(\mathbf{x}) = f(\mathbf{x})$ with $f(\mathbf{x})$ having an easy to evaluate analytical expression, see Sect. 4.1. This allows a good estimation of $Q^2_{\text{ideal}}(\mu)$ by $Q^2_{MC} = Q^2_{\text{ideal}}(\mu_M)$, see (1), where μ_M is the empirical measure for a large Monte-Carlo sample ($M = 10^6$), that will serve as reference when assessing the performance of each of the other estimators. We consider the validation designs built by FSSF, support points and kernel herding, presented in Sects. 3.1, 3.2, and 3.3, respectively, and, for each one, we compare the performances obtained for both the uniform and the weighted estimator of Sect. 2.2.

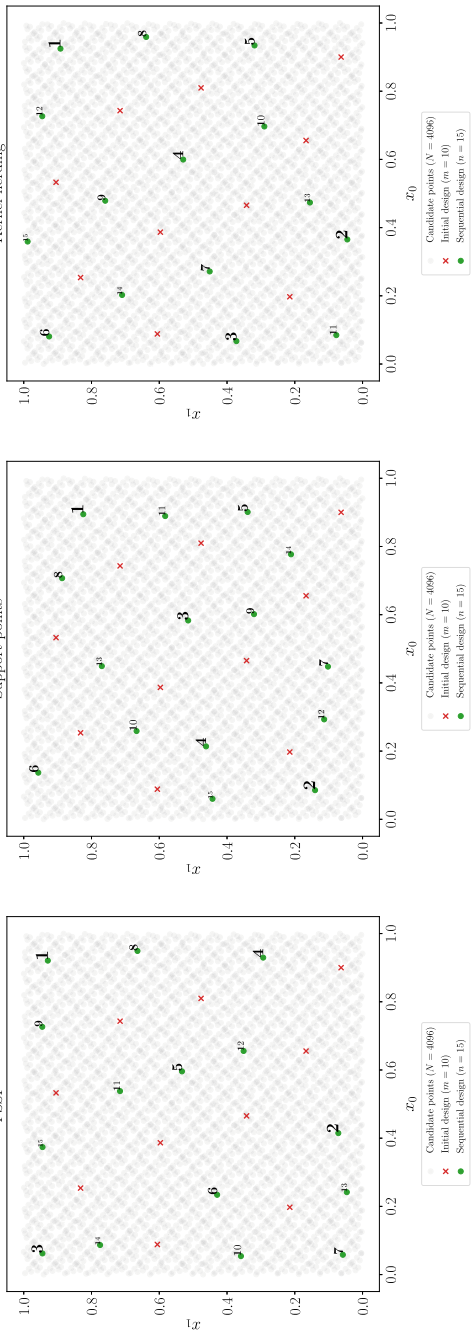


Fig. 1 Additional points (ordered, green) complementing an initial design (red crosses), μ is uniform on $[0, 1]$, the candidate points are in gray

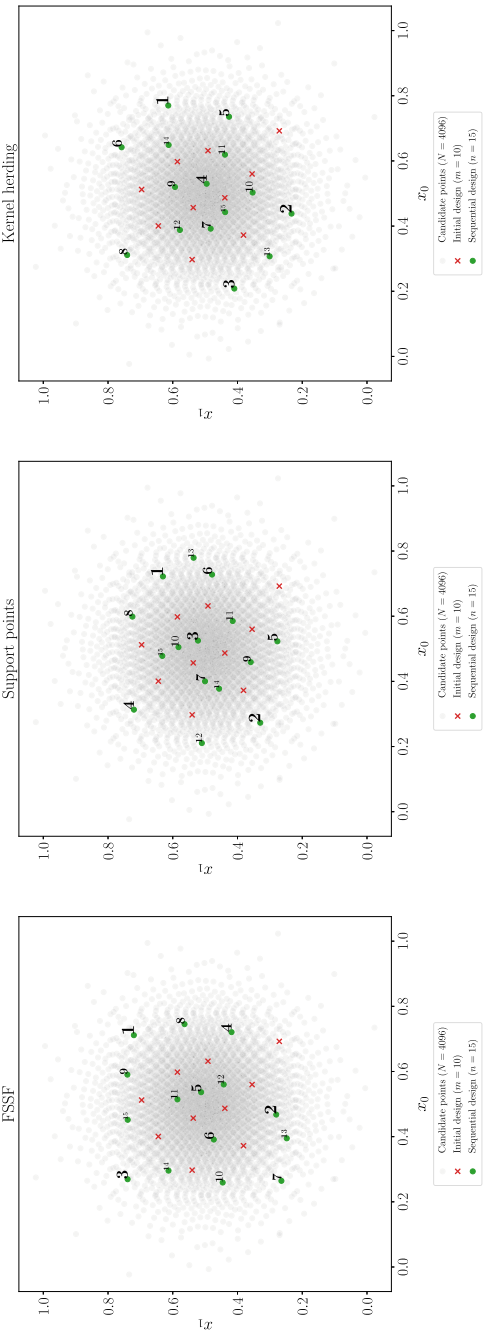


Fig. 2 Additional points (ordered, green) complementing an initial design (red crosses), μ normal, the candidate points are in gray

4.1 Test-cases

The training design $\mathbf{X}_m$ and the set $\mathcal{S}$ of potential test set points are as in Sect. 3.4. For test-cases 1 and 3, μ is the uniform measure on $\mathcal{X} = [0, 1]^d$, with $d = 2$ and $d = 8$, respectively; $\mathbf{X}_m$ is a maximin Latin hypercube design in $\mathcal{X}$, and $\mathcal{S}$ corresponds to the first N points $\mathbf{S}_N$ of Sobol' sequence in $\mathcal{X}$, complemented by the 2^d vertices. In the second test-case, $d = 2$, μ is the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, and the sets $\mathbf{X}_m$ and $\mathbf{S}_N$ must be transformed as explained in Sect. 3.1. There are $N = 2^{14}$ candidate points for test-cases 1 and 2 and $N = 2^{15}$ for test-case 3 (this value is rather moderate for a problem in dimension 8, but using a larger N yields numerical difficulties for support points; see Sect. 3.2).

For each test-case, a GP regression model is fitted to the m observations using ordinary Kriging [43] (a GP model with constant mean), with an anisotropic Matérn kernel with regularity parameter $5/2$: we substitute $[(\mathbf{x} - \mathbf{x}')^\top \mathbf{D}(\mathbf{x} - \mathbf{x}')]^{1/2}$ for $|x - x'|$ in (15), with $\mathbf{D}$ a diagonal matrix with diagonal elements $1/\theta_i^2$, and the correlation lengths θ_i are estimated by maximum likelihood via a truncated Newton algorithm. All calculations were done using the Python package OpenTURNS for uncertainty quantification [1]. The kernel used for kernel herding is different and corresponds to the tensor product of one-dimensional Matérn kernels (15), so that the potentials $P_{K,\mu}(\cdot)$ are known explicitly (see Appendix B); the correlations lengths are set to $\theta = 0.2$ in test-cases 1 and 3 ($d = 2$) and to $\theta = 0.7$ in test-case 3 ($d = 8$).

Assuming that a model is classified, in terms of the estimated value of its predictivity index Q^2 as "poor fitting" if $Q^2 \in [0.6, 0.8]$, "reasonably good fitting", when $Q^2 \in (0.8, 0.9]$, and "very good fitting" if $Q^2 > 0.9$, we selected, for each test-case three different sizes m of the training set such that the corresponding models cover all three possible situations. For all test-cases, the impact of the size n of the test set is studied in the range $n \in \{4, \ldots, 50\}$.

Test-case 1.

This test function is $f_1(\mathbf{x}) = h(2\,x_1 - 1, 2\,x_2 - 1)$, $(x_1, x_2) \in \mathcal{X} = [0, 1]^2$, with

$$h(u_1, u_2) = \frac{\exp(u_1)}{5} - \frac{u_2}{5} + \frac{u_2^6}{3} + 4u_2^4 - 4u_2^2 + \frac{7u_1^2}{10} + u_1^4 + \frac{3}{4u_1^2 + 4u_2^2 + 1}\,.$$

Color coded 3d and contour plots of f_1 for $\mathbf{X} \in \mathcal{X}$ are shown on the left panel of Fig. 3, showing that the function is rather smooth, even if its behaviour along the boundaries of $\mathcal{X}$, in particular close to the vertices, may present difficulties for some regression methods. The size of the training set for this function are: $m \in \{5, 15, 30\}$.

Test-case 2.

The second test function, plotted in the right panel of Fig. 3 for $\mathbf{x} \in [0, 1]^2$, is

$$f_2(\mathbf{x}) = \cos\left(5 + \frac{3}{2}x_1\right) + \sin\left(5 + \frac{3}{2}x_1\right) + \frac{1}{100}\left(5 + \frac{3}{2}x_1\right)\left(5 + \frac{3}{2}x_2\right)\,.$$

Training set sizes for this test-case are $m \in \{8, 15, 30\}$.

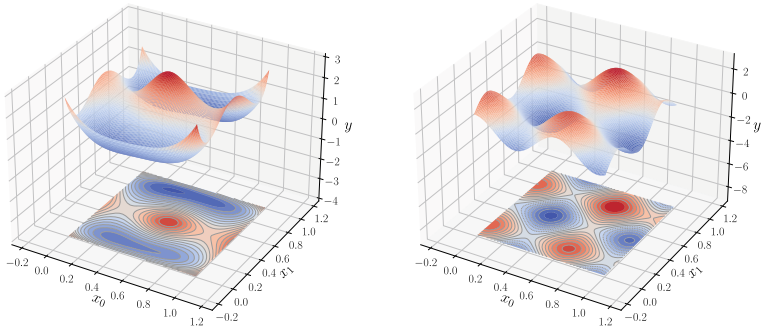


Fig. 3 Left: $f_1(\mathbf{x})$ (test-case 1); right: $f_2(\mathbf{x})$ (test-case 2); $\mathbf{x} \in \mathcal{X} = [0, 1]^2$

Test-case 3.
The third function is the so-called "gSobol" function, defined over $\mathcal{X} = [0, 1]^8$ by

$$f_3(\mathbf{x}) = \prod_{i=1}^{8} \frac{|4x_i - 2| + a_i}{1 + a_i}, \qquad a_i = i^2 .$$

This parametric function is very versatile as both the dimension of its input space and the coefficients a_i can be freely chosen. The sensitivity to input variables is determined by the a_i: the larger a_i is, the less f is sensitive to x_i. Larger training sets are considered for this test-case: $m \in \{15, 30, 100\}$.

4.2 Results and Analysis

The numerical results obtained in this section are presented in Figs. 4, 5, and 6. Each figure corresponds to one of the test-cases and gathers three sub-figures, corresponding to test sets with sizes m yielding poor (left), reasonably good (centre) or very good (right) fittings.

The baseline value of Q^2_{MC}, calculated with 10^6 Monte-Carlo points, is indicated by the black diamonds (the black horizontal lines). We assume that the error of Q^2_{MC} is much smaller than the errors of all other estimators, and compare the distinct methods through their ability to approximate Q^2_{MC}. For each sequence of nested test-sets ($n \in \{4, \ldots, 50\}$), the observed values of $\widehat{Q}^2_n$ (Eq. (2)) and Q^2_{n*} (equation (7)), are plotted as the solid and dashed lines, respectively.

The figures also show the value Q^2_{LOO} obtained by Leave-One-Out (LOO) cross validation, which is indicated at the left of each figure by a red diamond (values smaller than 0.25 are not shown). Note that, contrarily to the other methods considered, for LOO the test set is not disjoint from the training set, and thus the method does not satisfy the conditions set in the Introduction. As we repeat the

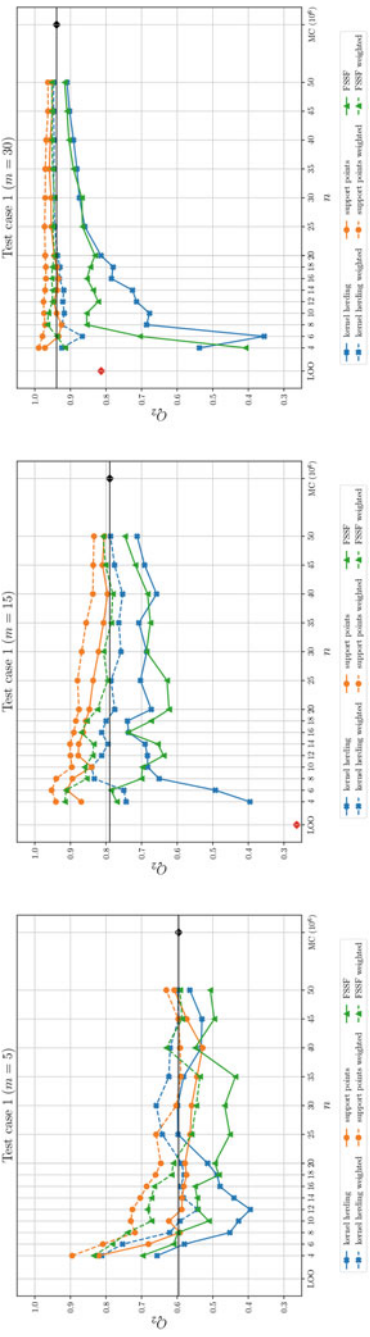


Fig. 4 Test-case 1: predictivity assessment of a poor (left), good (center) and very good (right) model with kernel herding, support points and FSSF test sets

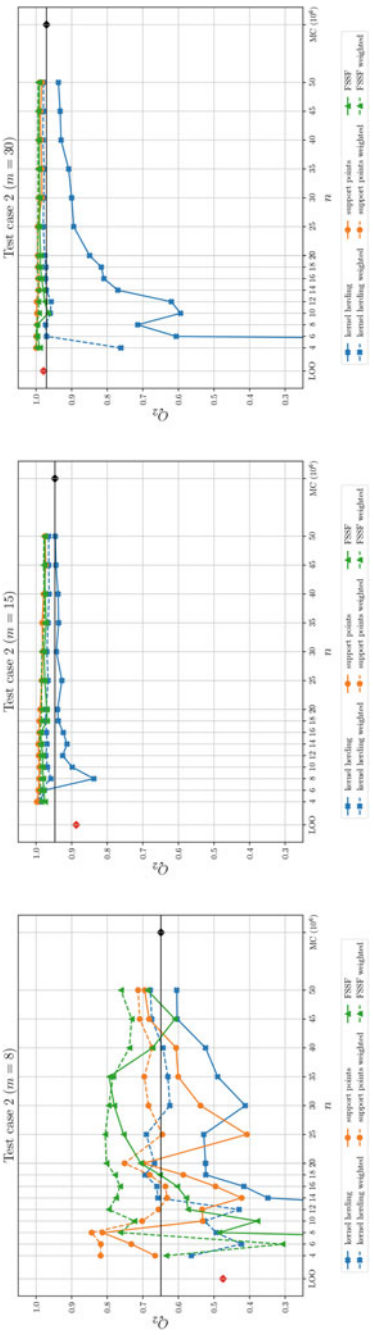


Fig. 5 Test-case 2: predictivity assessment of a poor (left), good (center) and very good (right) model with kernel herding, support points and FSSF test sets

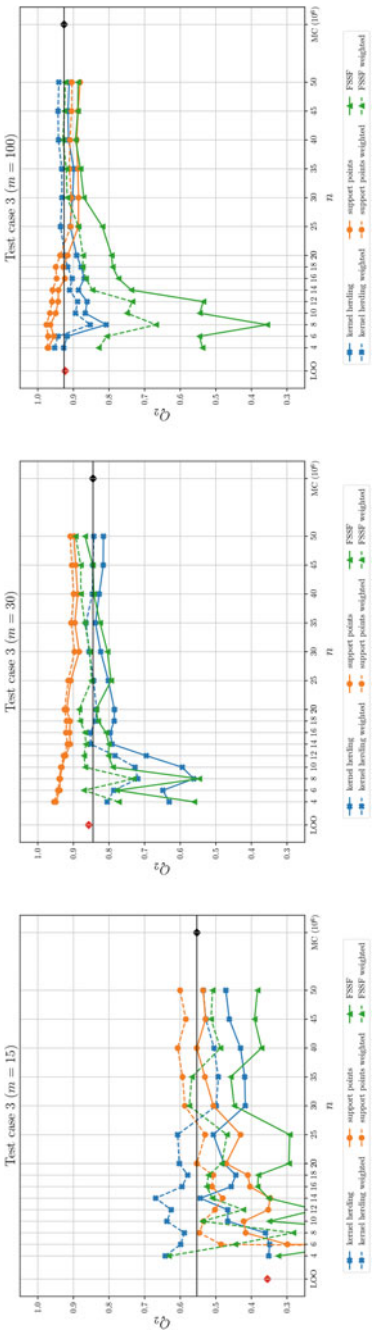


Fig. 6 Test-case 3: predictivity assessment of a poor (left), good (center) and very good (right) model with kernel herding, support points and FSSF test sets

complete model-fitting procedure for each training sample of size $m-1$, including the maximum-likelihood estimation of the correlation lengths of the Matérn kernel, the closed-form expressions of [13] cannot be used, making the computations rather intensive. As the three figures show, and as we should expect, Q^2_{LOO} tends to underestimate Q^2_{ideal}: by construction of the training set, LOO cross validation relies on model predictions at points $\mathbf{x}^{(i)}$ far from the other $m-1$ design points used to build the model, and thus tends to systematically overestimate the prediction error at $\mathbf{x}^{(i)}$. The underestimation of Q^2_{ideal} can be particularly severe when m is small, the training set being then necessarily sparse; see Fig. 4 where $Q^2_{LOO} < 0.3$ for $m = 5$ and 15.

Let us first concentrate on the non-weighted estimators (solid curves). We can see that the two MMD-based constructions, support points (in orange) and kernel herding (in blue), generally produce better validation designs than FSSF (green curves), leading to values of $\widehat{Q}^2_n$ that approach Q^2_{ideal} quicker as n increases. This is particularly noticeable for "good" and "very good" models (central and rightmost panels of all three figures). This supports the idea that test sets should complement the training set $\mathbf{X}_m$ by populating the holes it leaves in $\mathcal{X}$ while at the same time be able to mimic the target distribution μ, this second objective being more difficult to achieve for FSSF than for the MMD-based constructions.

Comparison of the two MMD based estimators reveals that support points tend to under-estimate ISE, leading to an over-confident assessment of the model predictivity, while kernel herding displays the expected behaviour, with a negative bias that decreases with n. The reason for the positive bias of estimates based on support points designs is not fully understood, but may be linked to the fact that support points tend to place validation points at "mid-range" from the designs (and not at the furthest points like FSSF or kernel herding), see central and rightmost panels in Fig. 1, and thus residuals at these points are themselves already better representatives of the local average errors.

We consider now the impact of the GP-based weighting of the residuals when estimating Q^2 (by Q^2_{n*}), which is related to the relative training-set/validation-set geometry (the manner in which the two designs are entangled in ambient space). The improvement resulting of applying residual weighting is apparent on all panels of the three figures, the dashed curves lying closer to Q^2_{ideal} than their solid counterparts; see in particular kernel herding (blue curve) in Fig. 4 and FSSF (green curve) in Fig. 5. Unexpectedly, the estimators based on support points seem to be rather insensitive to residual weighting, the dashed and solid orange curves being most of the time close to each other (and in any case, much closer that the green and blue ones). While the reason for this behavior deserves a deeper study, the fact that the support point designs—see Fig. 1—sample in a better manner the range of possible training-to-validation distances, being in some sense less space-filling than both FSSF and kernel herding, is again a plausible explanation for this weaker sensitivity to residual weighting.

Consider now comparison of the behaviour across test-cases. Setting aside the strikingly singular situation of test-case 2, for which kernel herding displays a pathological (bad) behaviour for the "very good" model, and all methods present an overall astonishing good behaviour, we can conclude that the details of the tested function

do not seem to play an important role concerning the relative merits of the estimators and validation designs.

We finally observe how the methods behave for models of distinct quality (m leading to poor, good or very good models), comparing the three panels in each figure. On the left panels, m is too small for the model η_m to be accurate, and all methods and test-set sizes are able to detect this. For models of practical interest (good and very good), the test sets generated with support points and kernel herding allow a reasonably accurate estimation of Q^2 with a few points. Note, incidentally, that except for test-case 2 (where the interplay with a non-uniform measure μ complicates the analysis), it is in general easier to estimate the quality of the very good model (right-most panel) than that of the good model (central panel), indicating that the expected complexity (the entropy) of the residual process should be a key factor determining how large the validation set must be. In particular, it may be that larger values of m allow for smaller values of n.

5 Numerical Results II: Splitting a Dataset into a Training Set and a Test Set

In this section, we illustrate the performance of the different designs and estimators considered in this paper when applied in the context of an industrial application, to split a given dataset of size N into training and test sets, with m and n points respectively, $m + n = N$. In contrast with [23], the observations $y(\mathbf{x}^{(i)})$, $i = 1, \ldots, N$, are not used in the splitting mechanism, meaning that it can be performed before the observations are collected and that there cannot be any selection bias related to observations (indeed, the use of observation values in a MMD-based splitting criterion may favour the allocation of the most different observations to different sets, training versus validation).

A ML model is fitted to the training data, and the data collected on the test-set are used to assess the predictivity of the model. The influence of the ratio $r_n = n/N = 1 - m/N$ on the quality assessment is investigated. We also consider Random Cross-Validation (RCV), where n points are chosen at random among the N points of the dataset: for each n, there are $\binom{N}{n}$ possible choices, and we randomly select $R = 1\,000$ designs among them. We fit a model to each of the m-point complementary designs ($m = N - n$), which yields an empirical distribution of Q^2 values for each ratio n/N considered.

5.1 Industrial Test-Case CATHARE

The test-case corresponds to the computer code CATHARE2 (for "Code Avancé de ThermoHydraulique pour les Accidents de Réacteurs à Eau"), which models

the thermal-hydraulic behavior inside nuclear pressurized water reactors [16]. The studied scenario simulates a hypothetical large-break loss of primary coolant accident for which the output of interest is the peak cladding temperature [11, 22]. The complexity of this application lies in the large run-time of the computer model (of the order of twenty minutes) and in the high dimension of the input space: the model involves 53 input parameters z_i, corresponding mostly to constants of physical laws, but also coding initial conditions, material properties and geometrical modeling. The z_i were independently sampled according to normal or log-normal distributions (see axes histograms in Fig. 7 corresponding to 10 inputs). These characteristics make this test-case challenging in terms of construction of a surrogate model and validation of its predictivity.

We have access to an existing Monte Carlo sample $\mathbf{Z}_N$ of $N = 1\,000$ points in $\mathbb{R}^{53}$, that corresponds to 53 independent random input configurations; see [22] for details. The output of the CATHARE2 code at these N points is also available. To reduce the dimensionality of this dataset, we first performed a sensitivity analysis [10] to eliminate inputs that do not impact the output significantly. This dimension-reduction step relies on the Hilbert-Schmidt Independence Criterion (HSIC), which is known as a powerful tool to perform input screening from a single sample of inputs and output values without reference to any specific ML regression model [9, 18]. HSIC-based statistical tests and their associated p-values are used to identify (with a 5%-threshold) inputs on which the output is significantly dependent (and therefore, also those of little influence). They were successfully applied to similar datasets from thermal-hydraulic applications in [30, 31]. The screened dataset only includes 10 influential inputs, over which the candidate set $\mathbf{X}_N$ used for the construction of the test-set $\mathbf{X}_n$ (and therefore of the complementary training set $\mathbf{X}_{N-n}$) is defined. An input-output scatter plot is presented in Fig. 7, showing that indeed the retained factors are correlated with the code output. The marginal distributions are shown as histograms along to the axes of the plots.

To include RCV in the methods to be compared, we need to be able to construct many (here, $R = 1\,000$) different models η_m for each considered design size m. Since Gaussian Process regression proved to be too expensive for this purpose, we settled for the comparatively cheaper Partial Least Squares (PLS) method [55], which retains acceptable accuracy. For each given training set, the model obtained is a sum of monomials in the 10 input variables. Note that models constructed with different training sets may involve different monomials and have different numbers of monomial terms.

5.2 Benchmark Results and Analysis

Figure 8 compares various ways of extracting an n-point test set from an N-point dataset to estimate model predictivity, for different splitting ratios $n/N \in \{0.1, 0.15, 0.2, \ldots, 0.9\}$.

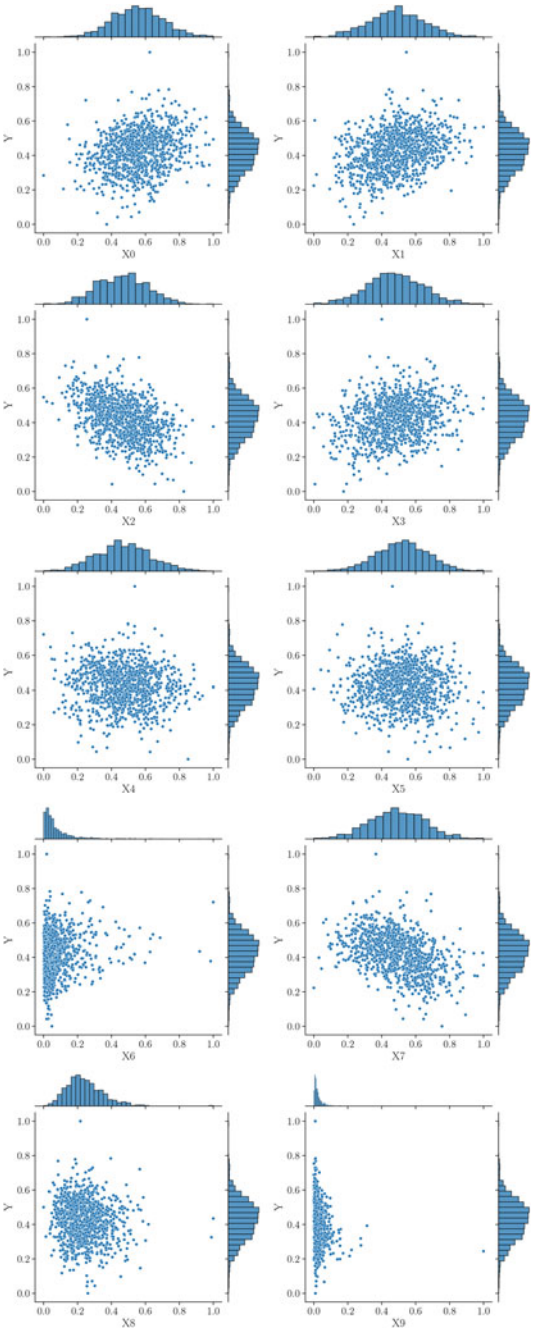


Fig. 7 Test-case CATHARE: inputs output scatter plots ($N = 10^3$)

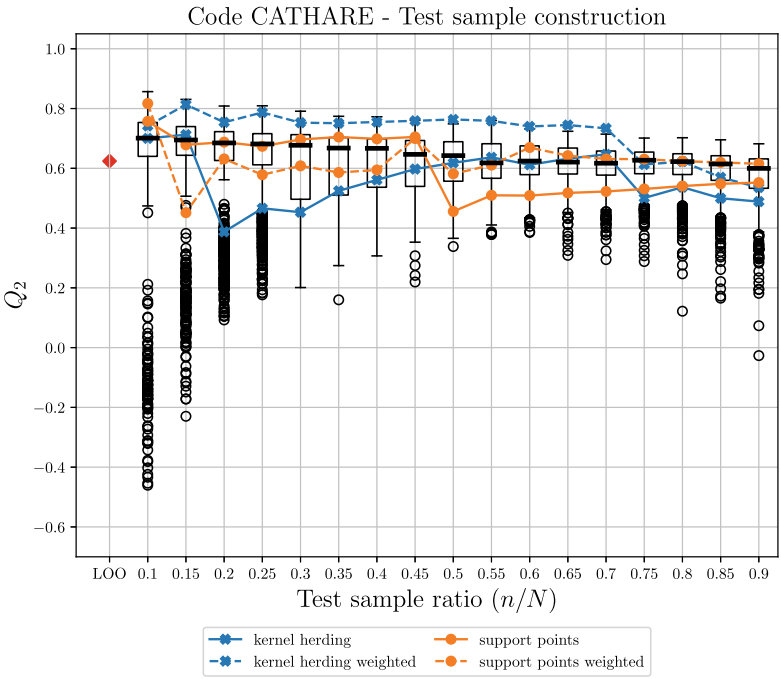


Fig. 8 Test-case CATHARE: estimated Q^2. The box-plots are for random cross-validation, the red diamond (left) is for Q^2_{LOO}

Consider RCV first. For each value of $r_n = n/N$, the empirical distribution of Q^2_{RCV} obtained from $R = 10^3$ random splittings of $\mathbf{X}_N$ into $\mathbf{X}_m \cup \mathbf{X}_n$ is summarized by a boxplot. Depending on r_n, we can roughly distinguish three behaviors. For $0.1 \leq r_n \lesssim 0.3$ the distribution is bi-modal, with the lower mode corresponding to unlucky test-set selections leading to poor performance evaluations. When $0.3 \lesssim n/N \lesssim 0.7$, the distribution looks uni-modal, revealing a more stable performance evaluation. Note that this is (partly) in line with the recommendations discussed in Sect. 1. For $r_n \gtrsim 0.7$, the variance of the distribution increases with r_n: many unlucky training sets lead to poor models. Note that the median of the empirical distribution slowly decreases as r_n increases, which is consistent with the intuition that the model predictivity should decrease when the size of the training set decreases.

For completeness, we also show by a red diamond on the left of Fig. 8 the value of Q^2_{LOO} computed by LOO cross-validation. In principle, being computed using the entire dataset, this value should establish an upper bound on the quality of models computed with smaller training sets. This is indeed the case for small training sets (rightmost values in the figure), for which the predictivity estimated by LOO is above the majority of the predictivity indexes calculated. But at the same time, we know that LOO cross-validation tends to overestimate the errors, which explains the higher predictivity estimated by some other methods when $m = N - n$ is large enough.

Compare now the behavior of the two MMD-based algorithms of Sect. 3, $\widehat{Q}_n^2$ (unweighted) and Q_{n*}^2 (weighted) are plotted using solid and dashed lines, respectively, for both kernel herding (in blue) and support points (in orange). FSSF test-sets are not considered, as the application of an iso-probabilistic transformation imposes knowledge of the input distribution, which is not known for this example. Compare first the unweighted versions of the two MMD-based estimators. For small values of the ratio r_n, $0.1 \lesssim r_n \lesssim 0.45$, the relative behavior of support points and kernel herding coincides with what we observed in the previous section, support points (solid orange line) estimating a better performance than kernel herding (solid blue line), which, moreover, is close to the median of the empirical distribution of Q_{RCV}^2. However, for $r_n \geq 0.5$, the dominance is reversed, support points estimating a worse performance than kernel herding.

As r_n increases up to $r_n \lesssim 0.7$ the solid orange and blue curves crossover, and it is now $\widehat{Q}_n^2$ for kernel herding that approximates the RCV empirical median, while the value obtained with support points underestimates the predictivity index. Also, note that for (irrealistic) very large values of r_n both support points and kernel herding estimate lower Q^2 values, which are smaller than the median of the RCV estimates.

Let us now focus on the effect of residual weighting, i.e., in estimators Q_{n*}^2 which use the weights computed by the method of Sect. 2.2, shown in dashed lines in Fig. 8. First, note that while for kernel herding weighting leads, as in the previous section, to higher estimates of the predictivity (compare solid and dashed blue lines), this is not the case for support points (solid and dashed orange curves), which, for small split ratios, produces smaller estimates when weighting is introduced. In the large r_n region, the behavior is consistent with what we saw previously, weighting inducing an increase of the estimated predictivity. It is remarkable—and rather surprising—that Q_{n*}^2 for support points (the dashed orange line) does not present the discontinuity of the uncorrected curve.

The sum $\sum_{i=1}^n w_i^*$ of the optimal weights of support points and kernel herding (6) is shown in Fig. 9 (orange and blue curves, respectively). The slow increase with n/N of the sum of kernel-herding weights (blue line) is consistent with the increase of the volume of the input region around each validation point when the size of the training set decreases. The behavior of the sum of weights is more difficult to interpret for support points (orange line) but is consistent with the behavior of Q_{n*}^2 on Fig. 8. Note that the energy-distance kernel (10) used for support points cannot be used for the weighting method of Sect. 2.2 as K_E is not positive definite but only conditionally positive definite. A full understanding of the observed curves would require a deeper analysis of the geometric characteristics of the designs generated by the two MMD methods, in particular of their interleaving with the training designs, which is not compatible with the space constraints of this manuscript.

While a number of unanswered points remain, in particular how deeply the behaviours observed may be affected by the poor predictivity resulting from the chosen PLS modeling methodology, the example presented in this section shows that the construction of test sets via MMD minimization and estimation of the predictivity index using the weighted estimator Q_{n*}^2 is promising as an efficient alternative to RCV: at a much lower computational cost, it builds performance estimates based on

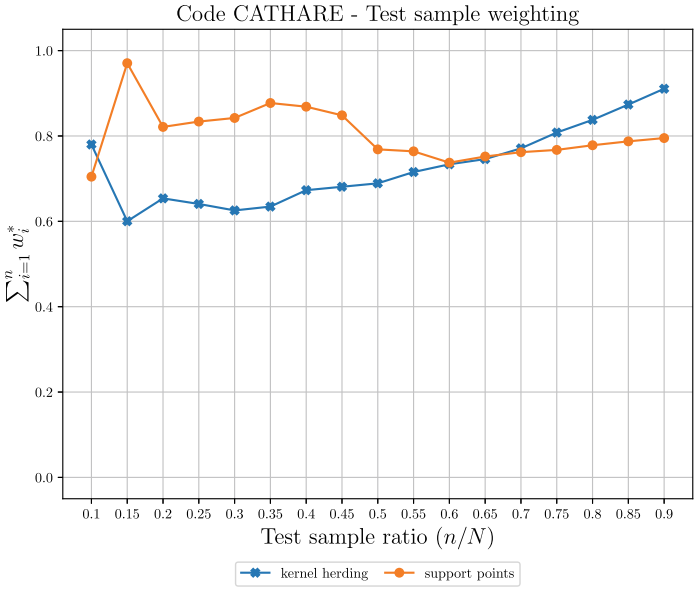


Fig. 9 Test-case CATHARE: sum of the weights (6)

independent data the model developers may not have access to. Moreover, kernel herding proved, in the examples studied in this manuscript, to be a more reliable option for designing the test set, exhibiting a behavior that is consistent with what is expected, and very good estimation quality when the residuals over the design points are appropriately weighted.

6 Conclusion

Our study shows that ideas and tools from the design of experiments framework can be transposed to the problem of test-set selection. This paper explored approaches based on support points, kernel herding and FSSF, considering the incremental construction of a test set (*i*) either as a particular space-filling design problem, where design points should populate the holes left in the design space by the training set, or (*i*) from the point of view of partitioning a given dataset into a training set and a test set.

A numerical benchmark has been performed for a panel of test-cases of different dimensions and complexity. Additionally to the usual predictivity coefficient, a new weighted metric (see [39]) has been proposed and shown to improve assessment of the predictivity of a given model for a given test set.

This weighting procedure appears very efficient for interpolators, like Gaussian process regression models, as it corrects the bias when the points in the test set used to predict the errors are far from the training points. For the first three test-cases

(Sect. 4), pairing one iterative design method with the weight-corrected estimator of the predictivity coefficient Q^2 shows promising results as the estimated Q^2 characteristic is close to the true one even for test-sets of moderate size.

Weighting can also be applied to models that do not interpolate the training data. For the industrial test-case of Sect. 5, the true Q^2 value is unknown, but the weight-corrected estimation Q^2_{n*} of Q^2 is close to the value estimated by Leave-One-Out cross validation and to the median of the empirical distribution of Q^2 values obtained by random k-fold cross-validation. At the same time, estimation by Q^2_{n*} involves a much smaller computational cost than cross-validation methods, and uses a dataset fully independent from the one used to construct the model.

To each of the design methods considered to select a test set a downside can be attached. FSSF requires knowledge of the input distribution to be able to apply an iso-probabilistic transformation if necessary; it tends to select many points along the boundary of the candidate set considered. Support points require the computation of the $N(N-1)/2$ distances between all pairs of candidate points, which implies important memory requirements for large N; the energy-distance kernel on which the method relies cannot be used for the weighting procedure. Finally, the efficient implementation of kernel herding relies on analytical expressions for the potentials $P_{K,\mu}$, see Appendices A and B, which are available for particular distributions (like the uniform and the normal) and kernels (like Matérn) only. The great freedom in the choice of the kernel K gives a lot of flexibility, but at the same time implies that some non-trivial decisions have to be made; also, the internal parameters of K, such as its correlation lengths, must to be specified. Future work should go beyond empirical rules of thumb and study the influence of these choices.

We have only computed numerical tests with independent inputs. Kernel herding and support points are both well suited for probability measures not being equal to the product of their marginals, which is a frequent case in real datasets. We have also only considered incremental constructions, as they allow to stop the validation procedure as soon as the estimation of the model predictivity is deemed sufficiently accurate, but it is also possible to select several points at once, using support points [29], or MMD minimization in general [54].

Further developments around this work could be as follows. Firstly, the incremental construction of a test set could be coupled with the definition of an appropriate stopping rule, in order to decide when it is necessary to continue improving the model (possibly by supplementing the initial design with the test set, which seems well suited to this). The MMD $d_{\overline{K}_{|m}}(\zeta_n^*, \mu)$ of Sect. 2.2 could play an important role in the derivation of such a rule. Secondly, the approach presented gives equal importance to all the d inputs. However, it seems that inputs with a negligible influence on the output should receive less attention when selecting a test set. A preliminary screening step that identifies the important inputs would allow the test-set selection algorithm to be applied on these variables only. For example, when a $\mathbf{X}_N \subset \mathbb{R}^d$ dataset is to be partitioned into $\mathbf{X}_m \cup \mathbf{X}_n$, one could use only $d' < d$ components to define the partition, but still use all d components to build the model and estimate its (weighted) Q^2. Note, however, that this would imply a slight violation of the condi-

tions mentioned in introduction, as it renders the test set dependent on the function observations.

Finally, in some cases the probability measure μ is known up to a normalizing constant. The use of a Stein kernel then makes the potential $P_{K,\mu}$ identically zero [4, 5], which would facilitate the application of kernel herding. Also, more complex problems involve functional inputs, like temporal signals or images, or categorical variables; the application of the methods presented to kernels specifically designed for such situations raises challenging issues.

Acknowledgements This work was supported by project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR). The authors are grateful to Guillaume Levillain and Thomas Bittar for their code development during their work at EDF. Thanks also to Sébastien Da Veiga for fruitful discussions.

Appendix

Appendix A: Maximum Mean Discrepancy

Let K be a positive definite kernel on $\mathcal{X} \times \mathcal{X}$, defining a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ of functions on $\mathcal{X}$, with scalar product $\langle f, g\rangle_{\mathcal{H}_K}$ and norm $\|f\|_{\mathcal{H}_K}$; see, e.g., [2]. For any $f \in \mathcal{H}_K$ and any probability measures μ and ξ on $\mathcal{X}$, we have

$$\left|\int_{\mathcal{X}} f(\mathbf{x})\,\mathrm{d}\xi(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})\right| = \left|\int_{\mathcal{X}} \langle f, K_{\mathbf{x}}\rangle_{\mathcal{H}_K}\,\mathrm{d}(\xi-\mu)(\mathbf{x})\right| = \left|\langle f, (P_{K,\xi} - P_{K,\mu}\rangle_{\mathcal{H}_K}\right|, \tag{16}$$

where we have denoted $K_{\mathbf{x}}(\cdot) = K(\mathbf{x}, \cdot)$ and used the reproducing property $f(\mathbf{x}) = \langle f, K_{\mathbf{x}}\rangle_{\mathcal{H}_K}$ for all $\mathbf{x} \in \mathcal{X}$, and where, for any probability measure ν on $\mathcal{X}$ and $\mathbf{x} \in \mathcal{X}$,

$$P_{K,\nu}(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}')\,\mathrm{d}\nu(\mathbf{x}'), \tag{17}$$

is the potential of ν at $\mathbf{x}$. $P_{K,\nu} \in \mathcal{H}_K$ and is called kernel embedding of ν in ML. In some cases, the potential can be expressed analytically (see. Appendix 6), otherwise it can be estimated by numerical quadrature (Quasi Monte Carlo). Cauchy-Schwartz inequality applied to (16) gives

$$\left|\int_{\mathcal{X}} f(\mathbf{x})\,\mathrm{d}\xi(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})\right| \leq \|f\|_{\mathcal{H}_K}\,\|P_{K,\xi} - P_{K,\mu}\|_{\mathcal{H}_K}$$

and therefore

$$\|P_{K,\xi} - P_{K,\mu}\|_{\mathcal{H}_K} = \sup_{f \in \mathcal{H}_K:\ \|f\|_{\mathcal{H}_K}=1} \left| \int_{\mathcal{X}} f(\mathbf{x})\, \mathrm{d}\xi(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x}) \right| .$$

The Maximum Mean Discrepancy (MMD) between ξ and μ (for the kernel K and set $\mathcal{X}$) is $d_K(\xi, \mu) = \|P_{K,\xi} - P_{K,\mu}\|_{\mathcal{H}_K}$. Direct calculation gives

$$d_K^2(\xi, \mu) = \|P_{K,\xi} - P_{K,\mu}\|_{\mathcal{H}_K}^2 = \int_{\mathcal{X}^2} K(\mathbf{x}, \mathbf{x}')\, \mathrm{d}(\xi - \mu)(\mathbf{x}) \mathrm{d}(\xi - \mu)(\mathbf{x}') \quad (18)$$

$$= \mathbb{E}_{\zeta,\zeta' \sim \xi} K(\zeta, \zeta') + \mathbb{E}_{\zeta,\zeta' \sim \mu} K(\zeta, \zeta') - 2\mathbb{E}_{\zeta \sim \xi, \zeta' \sim \mu} K(\zeta, \zeta') , \quad (19)$$

where the random variables ζ and ζ' in (19) are independent, see [49]. When K is the energy distance kernel (10), one recovers the expression (11) for the corresponding MMD. One may refer to [51] for an illuminating exposition on MMD, kernel embedding, and conditions on K (the notion of characteristic kernel) that make d_K a metric on the space of probability measures on $\mathcal{X}$. The distance and Matérn kernels considered in this paper are characteristic.

Appendix B: Analytical Computation of Potentials for Matérn Kernels

As for tensor-product kernels, the potential is the product of the one-dimensional potentials, we only consider one-dimensional input spaces.

For μ the uniform distribution on $[0, 1]$ and K the Matérn kernel $K_{5/2,\theta}$ with smoothness $\nu = 5/2$ and correlation length θ, see (15), we get

$$P_{K_{5/2,\theta},\mu}(x) = \frac{16\theta}{3\sqrt{5}} - \frac{1}{15\theta}(S_\theta(x) + S_\theta(1 - x)),$$

where

$$S_\theta(x) = \exp\left(-\frac{\sqrt{5}}{\theta}x\right)\left(5\sqrt{5}x^2 + 25\theta x + 8\sqrt{5}\theta^2\right).$$

The expressions $P_{K_{\nu,\theta},\mu}(x)$ for $\nu = 1/2$ and $\nu = 3/2$ can be found in [40].

When μ is the standard normal distribution $\mathcal{N}(0, 1)$, the potential $P_{K_{5/2,\theta},\mathcal{N}(0,1)}$ is $P_{K_{5/2,\theta},\mathcal{N}(0,1)}(x) = T_\theta(x) + T_\theta(-x)$, where

$$T_\theta(x) = \frac{1}{6}\left(\frac{5}{\theta^2}x^2 + \left(3 - \frac{10}{\theta^2}\right)\frac{\sqrt{5}}{\theta}x + \frac{5}{\theta^2}\left(\frac{5}{\theta^2} - 2\right) + 3\right)$$
$$\times \operatorname{erfc}\left(\frac{\frac{\sqrt{5}}{\theta} - x}{\sqrt{2}}\right)\exp\left(\frac{5}{2\theta^2} - \frac{\sqrt{5}}{\theta}x\right)$$
$$+\frac{1}{3\sqrt{2\pi}}\frac{\sqrt{5}}{\theta}\left(3 - \frac{5}{\theta^2}\right)\exp\left(-\frac{x^2}{2}\right).$$

References

1. Baudin, M., Dutfoy, A., Iooss, B., Popelin, A-P.: Open TURNS: An industrial software for uncertainty quantification in simulation. In: Ghanem, R., Higdon, D., Owhadi, H. (eds.) Springer Handbook on Uncertainty Quantification, pp. 2001–2038. Springer (2017)
2. Berlinet , A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer (2004)
3. Borovicka, T., Jr. Jirina, M., Kordik, P., Jirina, M.: Selecting representative data sets. In: Karahoca, A. (eds) Advances in Data Mining, Knowledge Discovery and Applications, pp. 43–70. INTECH (2012)
4. Chen, W.Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., Oates, C.: Stein Point Markov Chain Monte Carlo. arXiv preprint. arXiv:1905.03673 (2019)
5. Chen, W.Y., Mackey, L., Gorham, J., Briol, F.-X., Oates, C.J.: Stein Points. Proc. ICML (2018). arXiv preprint arXiv:1803.10161v4
6. Chen, Y., Welling, M., Smola, A.: Super-samples from kernel herding. In Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, pp. 109–116. AUAI Press (2010)
7. Chevalier, C., Bect, J., Ginsbourger, D., Picheny, V., Richet, Y., Vazquez, E.: Fast kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics **56**, 455–465 (2014)
8. Crombecq, K., Laermans, E., Dhaene, T.: Efficient space-filling and non-collapsing sequential design strategies for simulation-based modelling. Eur. J. Oper. Res. **214**, 683–696 (2011)
9. Da Veiga, S.: Global sensitivity analysis with dependence measures. J. Stat. Comput. Simul. **85**, 1283–1305 (2015)
10. Da Veiga, S., Gamboa, F., Iooss, B., Prieur, C.: Basics and Trends in Sensitivity Analysis. Theory and Practice in R. SIAM (2021)
11. de Crécy, A., Bazin, P., Glaeser, H., Skorek, T., Joufcla, J., Probst, P., Fujioka, K., Chung, B.D., Oh, D.Y., Kyncl, M., Pernica, R., Macek, J., Meca, R., Macian, R., D'Auria, F., Petruzzi, A., Batet, L., Perez, M., Reventos, F.: Uncertainty and sensitivity analysis of the LOFT L2–5 test: results of the BEMUSE programme. Nucl. Eng. Design **12**, 3561–3578 (2008)
12. Demay, C., Iooss, B., Le Gratiet, L., Marrel, A.: Model selection for Gaussian Process regression: an application with highlights on the model variance validation. Qual. Reliab. Eng. Int. J. **38**, 1482–1500 (2022). https://doi.org/10.1002/qre.2973
13. Dubrule, O.: Cross validation of kriging in a unique neighborhood. J. Int. Assoc. Math. Geol. **15**(6), 687–699 (1983)
14. ENIQ: Qualification of an AI/ML NDT system—Technical basis. NUGENIA, ENIQ Technical Report (2019)
15. Fang, K.-T., Li, R., Sudjianto, A.: Design and Modeling for Computer Experiments. Chapman & Hall/CRC (2006)
16. Geffraye, G., Antoni, O., Farvacque, M., Kadri, D., Lavialle, G., Rameau, B., Ruby, A.: CATHARE2 V2.5_2: a single version for various applications. Nucl. Eng. Des. **241**, 4456–4463 (2011)

17. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016)
18. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings Algorithmic Learning Theory, pp. 63–77. Springer-Verlag (2005)
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer (2009)
20. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., Habli, I.: Guidance on the assurance of machine learning in autonomous systems (AMLAS). University of York, Assuring Autonomy International Programme (AAIP) (2021)
21. Iooss, B.: Sample selection from a given dataset to validate machine learning models. In Proceedings of 50th Meeting of the Italian Statistical Society (SIS2021), pp. 88–93. Pisa, Italy, June (2021)
22. Iooss, B., Boussouf, L., Feuillard, V., Marrel, A.: Numerical studies of the metamodel fitting and validation processes. Int. J. Adv. Syst. Measure. **3**, 11–21 (2010)
23. Joseph, V.R., Vakayil, A.: SPlit: an optimal method for data splitting. Technometrics **64**(2), 166–176 (2022)
24. Kennard, R.W., Stone, L.A.: Computer aided design of experiments. Technometrics **11**, 137–148 (1969)
25. Kleijnen, J.P.C., Sargent, R.G.: A methodology for fitting and validating metamodels in simulation. Eur. J. Oper. Res. **120**, 14–29 (2000)
26. Lemaire, M., Chateauneuf, A., Mitteau, J.-C.: Structural Reliability. Wiley (2009)
27. Li, W., Lu, L., Xie, X., Yang, M.: A novel extension algorithm for optimized Latin hypercube sampling. J. Stat. Comput. Simul. **87**, 2549–2559 (2017)
28. Lorenzo, G., Zanocco, P., Giménez, M., Marquès, M., Iooss, B., Bolado-Lavin, R., Pierro, F., Galassi, G., D'Auria, F., Burgazzi, L.: Assessment of an isolation condenser of an integral reactor in view of uncertainties in engineering parameters. Sci. Technol. Nucl. Install. (2011). https://doi.org/10.1155/2011/827354
29. Mak, S., Joseph, V.R.: Support points. Ann. Stat. **46**, 2562–2592 (2018)
30. Marrel, A., Chabridon, V.: Statistical developments for target and conditional sensitivity analysis: Application on safety studies for nuclear reactor. Reliab. Eng. Syst. Saf. **214**, 107711 (2021)
31. Marrel, A., Iooss, B., Chabridon, V.: The ICSCREAM methodology: identification of penalizing configurations in computer experiments using screening and metamodel - Applications in thermal-hydraulics. Nucl. Sci. Eng. **196**, 301–321 (2022). https://doi.org/10.1080/00295639.2021.1980362
32. Molnar, C.: Interpretable Machine Learning. github (2019)
33. Morris, M.D., Mitchell, T.J.: Exploratory designs for computational experiments. J. Stat. Planning Inference **43**, 381–402 (1995)
34. Müller, W.G.: Collecting Spatial Data, 3rd edn. Springer (2007)
35. Nash, J., Sutcliffe, J.: River flow forecasting through conceptual models part I-A discussion of principles. J. Hydrol. **10**(3), 282–290 (1970)
36. Nogales Gómez, A., Pronzato, L., Rendas, M.-J.: Incremental space-filling design based on coverings and spacings: improving upon low discrepancy sequences. J. Stat. Theory Pract. **15**(4), 77 (2021)
37. Pronzato, L.: Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. Statistics and Computing, to appear (2022), hal-03114891. arXiv:2101.07564
38. Pronzato, L., Müller, W.: Design of computer experiments: space filling and beyond. Stat. Comput. **22**, 681–701 (2012)
39. Pronzato, L., Rendas, M.-J.: Validation design I: construction of validation designs via kernel herding. Preprint (2021), hal-03474805. arXiv:2112.05583
40. Pronzato, L., Zhigljavsky, A.A.: Bayesian quadrature and energy minimization for space-filling design. SIAM/ASA J. Uncertainty Quant. **8**, 959–1011 (2020)
41. Qian, P.Z.G., Ai, M., Wu, C.F.J.: Construction of nested space-filling designs. Ann. Stat. **37**, 3616–3643 (2009)

42. Qian, P.Z.G., Wu, C.F.J.: Sliced space filling designs. Biometrika **96**, 945–956 (2009)
43. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
44. Santner, T., Williams, B., Notz, W.: The Design and Analysis of Computer Experiments. Springer (2003)
45. Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K.: Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann. Stat. **41**(5), 2263–2291 (2013)
46. Shang, B., Apley, D.W.: Fully-sequential space-filling design algorithms for computer experiments. J. Qual. Technol. **53**(2), 173–196 (2021)
47. Sheikholeslami, R., Razavi, S.: Progressive Latin hypercube sampling: an efficient approach for robust sampling-based analysis of environmental models. Environ. Model. Softw. **93**, 109–126 (2017)
48. Smith, R.C.: Uncertainty Quantification. SIAM (2014)
49. Smola, A., Gretton, A., Song, L., Schölkopf, B.: A Hilbert space embedding for distributions. In International Conference on Algorithmic Learning Theory, pp. 13–31. Springer (2007)
50. Snee, R.D.: Validation of regression models: methods and examples. Technometrics **19**, 415–428 (1977)
51. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. J. Mach. Learn. Res. **11**, 1517–1561 (2010)
52. Székely, G.J., Rizzo, M.L.: Testing for equal distributions in high dimension. InterStat **5**, 1–6 (2004)
53. Székely, G.J., Rizzo, M.L.: Energy statistics: a class of statistics based on distances. J. Stat. Planning Inference **143**, 1249–1272 (2013)
54. Teymur, O., Gorham, J., Riabiz, M., Oates, C.J.: Optimal quantisation of probability measures using maximum mean discrepancy. In International Conference on Artificial Intelligence and Statistics, pp. 1027–1035 (2021). arXiv preprint arXiv:2010.07064v1
55. Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. **58**(2), 109–130 (2001)
56. Xu, Y., Goodacre, R.: On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Testing **2**, 249–262 (2018)

Multiple Measures Realized GARCH Models

Antonio Naimoli and Giuseppe Storti

Abstract Realized volatility has become the most popular empirical measure in fitting and forecasting volatility. However, as the properties of this class of estimators depend on the sampling frequency of intraday returns, a number of alternative realized estimators have been proposed, generating additional uncertainty in the modelling process. Aiming to mitigate the impact of modelling uncertainty in forecasting tail-risk, this paper investigates the benefits of combining information from several realized measures computed at multiple frequencies. In this framework, extensions of the Realized GARCH model based both on feature selection methods and time-varying parameters are proposed. To assess the implications for financial risk management, an application to the prediction of Value-at-Risk and Expected Shortfall for the Standard & Poor's 500 Index is presented. We find that significant forecasting gains result from modelling approaches combining several realized multi-frequency measures.

Keywords Realized GARCH · Realized volatility measures · Sampling frequency · PCA · Tail risk forecasting

1 Introduction

High-frequency data analysis has continuously gained importance in recent years as the demand for high-quality intraday data is growing at the same time, both for decision making and research purposes. The use of realized volatility (RV) measures to accurately estimate the integrated variance of a price process over some interval of time is currently one of the most widely used approaches for modelling and forecasting volatility. RV measures rely on high-frequency prices to estimate the daily

A. Naimoli (✉) · G. Storti
Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES),
Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy
e-mail: anaimoli@unisa.it

G. Storti
e-mail: storti@unisa.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_21

volatility of financial assets. If prices were not affected by market microstructure noise, then the RV, i.e. the sum of squared high-frequency intra-day returns, would provide a consistent estimate of volatility [7]. This suggests that RV should be calculated using tick-by-tick data or intra-day returns sampled at the highest possible frequency.

However, the presence of market microstructure noise in high-frequency data affects the properties of the realized measures making the estimation of the returns' variance more challenging. Consequently, a bias/variance trade-off induced by the presence of microstructure noise characterizes the RV, which suffers from a bias problem when the sampling frequency of intra-day returns tends to increase (see, among others, [3, 9, 27]). In this framework, kernel-based estimators [10, 27], subsample-based estimators [3, 37] and jump-robust estimators [8, 13] can be used to estimate the quadratic variation of an underlying efficient price process from high-frequency noisy data [11].

In addition, the degree of bias due to noise and jumps could be inherently time-varying Aït-Sahalia and Yu [4]. Accordingly, this implies that when selecting the optimal frequency for calculating realized estimators, the solution to the problem of identifying the optimal trade-off between bias and variance could be time-varying. In this context, one solution might be to use an adaptive frequency, which varies over time depending on market conditions.

This discussion clearly highlights how the choice of the realized measure and the sampling frequency of intra-day returns can significantly influence volatility and tail risk forecasts. The aim of this paper is to assess the gains for tail risk forecasting of combining multiple realized volatility measures. This is done by employing different volatility estimators characterized by distinct sensitivity to noise and jumps, mixing realized measures at different sampling frequencies and controlling for the time-varying bias/variance trade-off in volatility estimation. Therefore, the idea is to adaptively merge the information from a wide set of realized measures in order to improve the accuracy of Value-at-Risk (VaR) and Expected Shortfall (ES) forecasts. To this purpose, extensions of the Realized GARCH model by Hansen et al. [26] are proposed.

In particular, the first extension allows to assign time-varying weights to the realized measures involved in the model specification through a weighting function whose dynamics are driven by an appropriately chosen state variable. Namely, future volatility forecasts are driven by the variation of a weighted average of the realized measures involved, where the weights vary over time (on a daily scale) depending on the estimated amount of noise and jumps. The focus here is on two realized volatility measures based on different discretization grids, so the proposed specification is called *Adaptive Frequency* Realized GARCH (AF-RGARCH) model. Models in this class include distinct measurement equations for each of the realized measures considered. We also develop a parsimonious version of the AF-RGARCH characterized by a single measurement equation, called the *Single* equation AF-RGARCH (SAF-RGARCH) model. However, although the AF-RGARCH can be, in principle, easily generalized to incorporate more than two measures, in practice this possibility is limited by the rapid increase in the number of parameters to be estimated. In

addition, the predictive performance of the model critically depends on the specific pair of realized measures of interest, but a rigorous procedure for their identification is missing.

To overcome the problem of having to choose the *right* pair of realized measures, we propose an alternative modelling approach that allows to jointly use a wide range of volatility measures. Making use of high-frequency data, since Andersen and Bollerslev [6], several estimators have been proposed in the literature to estimate the integrated variance of the price process. In order to parsimoniously deal with the information from such a large set of volatility estimators, a factor-type structure on the dynamic volatility update equation is imposed. That is, to identify the set of latent factors, we resort to the use of standard dimension reduction techniques such as the Principal Component Analysis (PCA), which returns unconditionally uncorrelated factors and has already been applied in the literature on Multivariate Factor GARCH models [5, 19, 34]. Therefore, the resulting model is called a PC-RGARCH.

An application to the S&P 500 Index provides evidence that modelling approaches based on the combination of different frequencies and estimation formulas lead to significant gains in the accuracy of tail risk forecasts, as revealed by the Model Confidence Set [28]. That is, the out-of-sample results show that the lowest losses for both VaR and joint (VaR,ES) forecasts, at all risk levels considered, are obtained from model specifications that combine several realized multi-frequency measures.

The paper is structured as follows: in Sect. 2 we briefly review the Realized GARCH model by Hansen et al. [26]. The class of Adaptive Frequency RGARCH models is introduced in Sect. 3, while the PC-RGARCH is presented in Sect. 4. Section 5 provides details on the estimation procedures used to fit the proposed models. Sections from 6 to 7 focus on the empirical application: Sect. 6 describes the dataset used and Sect. 7 focuses on the out-of-sample tail-risk forecasting analysis. Finally, Sect. 8 concludes.

2 Realized GARCH Models

The Realized GARCH (RGARCH) model of Hansen et al. [26] introduces a flexible framework for jointly modelling returns and realized volatility measures. Differently from the standard GARCH, the RGARCH relates the observed realized measure to latent volatility through a measurement equation, also including an asymmetric response to shocks, thus making the model very flexible and dynamically complete, allowing the generation of multi-step forecasts. The RGARCH model (with log specification) is given by

$$r_t = \sqrt{h_t}\, z_t \tag{1}$$

$$\log(h_t) = \omega + \beta \log(h_{t-1}) + \gamma \log(x_{t-1}) \tag{2}$$

$$\log(x_t) = \xi + \varphi \log(h_t) + \tau(z_t) + u_t \tag{3}$$

where r_t is the log-return for day t and x_t is a realized measure. Also, $\mathcal{F}_{t-1}$ denotes the information set at time $t-1$, then $h_t = var(r_t|\mathcal{F}_{t-1})$, with $z_t \overset{iid}{\sim} (0, 1)$ and $u_t \overset{iid}{\sim} (0, \sigma_u^2)$ mutually independent.

The three equations characterizing the RGARCH are, the return equation (1), the volatility equation (2) and the measurement equation (3), respectively. The measurement equation is designed to capture the contemporaneous dependence between latent volatility and the realized measure, where the term $\tau(z_t) = \tau_1 z_t + \tau_2(z_t^2 - 1)$ is used to model a leverage-type effect.

Substituting the measurement equation into the volatility equation leads to

$$\log(h_t) = (\omega + \xi\gamma) + (\beta + \varphi\gamma)\log(h_{t-1}) + \gamma\, w_{t-1}, \tag{4}$$

where $w_t = \tau(z_t) + u_t$ and $E(w_t) = 0$, with the restriction $(\beta + \varphi\gamma) < 1$ ensuring strict stationarity of the RGARCH(1,1) process [26, 29].

Hansen et al. [26] assume Gaussian errors for z_t and u_t, although other assumptions on the conditional distribution of returns can be made (see, e.g., [17, 22, 35]).

3 Adaptive Frequency Realized GARCH Models

This section presents extensions of the standard RGARCH model that allows to exploit information from realized volatility measures based on different sampling frequencies. The proposed model evolves over the RGARCH specification under two different aspects. First, volatility dynamics are given by a weighted average of two realized volatility measures, built from returns sampled at different frequencies, where the weights are time-varying and adaptively determined. Second, as in Hansen and Huang [25], the model includes a different measurement equation for each of the realized measures considered. Due to its inherent features, the resulting model specification is called the Adaptive Frequency Realized GARCH (AF-RGARCH) model.

Namely, the AF-RGARCH model is defined by the following equations

$$\log(h_t) = \omega + \beta\, \log(h_{t-1}) + \gamma\, \log(\tilde{x}_{t-1}) \tag{5}$$

$$\log(\tilde{x}_t) = \lambda_t\, \log\left(x_t^{(H)}\right) + (1 - \lambda_t)\, \log\left(x_t^{(L)}\right) \tag{6}$$

$$\log\left(x_t^{(j)}\right) = \xi_j + \varphi_j\, \log(h_t) + \tau_j(z_t) + u_{j,t} \qquad j = H, L \tag{7}$$

where $x_t^{(H)}$ and $x_t^{(L)}$ denote realized measures computed at a higher and a lower frequency, respectively.

Accordingly, $\tilde{x}_t$ is a weighted average of two realized measures based on different sampling frequencies, where the time-varying weight λ_t is of the form

$$\lambda_t = \lambda_0 + \lambda_1 \log\left(R_t^{(H,L)}\right)\mathcal{I}_{\left(R_t^{(H,L)}\leq 1\right)} + \lambda_2 \log\left(R_t^{(H,L)}\right)\mathcal{I}_{\left(R_t^{(H,L)}>1\right)}, \qquad (8)$$

where $R_t^{(H,L)} = RQ_t^{(H)}/RQ_t^{(L)}$ and $RQ_t^{(j)}$ the Realized Quarticity (RQ) at frequency j

$$RQ_t^{(j)} = \frac{M^{(j)}}{3}\sum_{i=1}^{M^{(j)}}\left(r_{t,i}^{(j)}\right)^4, \qquad j = H, L.$$

The function $\mathcal{I}_{(\cdot)}$ takes the value 1 if the given condition occurs and 0 otherwise, while $r_{t,i}^{(j)}$ is the j-frequency log-return observed on the i-th intra-day sub-interval and $M^{(j)}$ is the number of sampled intra-day intervals at the same frequency.

This allows the conditional variance dynamics of an AF-RGARCH model to be driven by an "artificial" realized measure constructed as a weighted average of $x^{(H)}$ and $x^{(L)}$, with weights varying over time as a function of the state variable $R_t^{(H,L)}$. The use of the log-transformed ratio between realized quarticities computed at high and low frequencies in λ_t is motivated by the fact that the RQ highlights the sensitivity of the fourth moment of intra-day returns to outliers and market microstructure noise. However, these effects tend to wane as the sampling frequency of returns shrinks [9, 12]. Therefore, $R_t^{(H,L)}$ has the role of correcting for upward and downward biases that might affect the realized measures, characterizing the effects of negative and positive biases differently. Namely, λ_t is modelled as a piece-wise linear function of $\log\left(R_t^{(H,L)}\right)$, assigning potentially different impacts to values of $R_t^{(H,L)}$ lower and greater than 1, respectively.

In a noise-free world, $x^{(H)}$ is more efficient than $x^{(L)}$. However, the presence of microstructure noise and outliers could introduce a bias component into the realized $x^{(H)}$ estimator. Thus, the higher frequency estimator should be used whenever the impact of noise is negligible but, at the same time, in order to correct for the bias, we would like to down-weight the $x^{(H)}$ estimator whenever this impact becomes more substantial.

When $R_t^{(H,L)} \approx 1$, the value of λ_t, that is the weight given to the high frequency component, is $\approx \lambda_0$. On the other hand, values of $R_t^{(H,L)} \neq 1$ will lead to deviations of λ_t from this long-run level. For example, if λ_2 is negative, the value λ_t will be proportionally reduced whenever $R_t^{(H,L)} > 1$.

In this framework, the role of the AF-RGARCH model is to provide a solution for identifying the optimal bias/variance trade-off in volatility forecasting.

Remark 1 (constant weights model). When λ_1 and λ_2 are (jointly) not statistically significant, the data provide evidence that the bias correction term $R^{(H,L)}t$ is unnecessary. It follows that, $log(\tilde{x}_t)$ is given by the constant weights convex combination

$$\log(\tilde{x}_t) = \lambda_0 \log\left(x_t^{(H)}\right) + (1-\lambda_0)\log\left(x_t^{(L)}\right). \qquad (9)$$

Therefore, in this case, λ_0 and $(1-\lambda_0)$ determine, respectively, the importance assigned to $x_t^{(H)}$ and $x_t^{(L)}$ for the overall realized measure $\tilde{x}_t$. Also, for $\lambda_1 = \lambda_2 = 0$ and λ_0 equal to either 0 or 1, the standard RGARCH model is obtained as a special case.

Remark 2 (Time-varying AR(1) representation). The AF-RGARCH model can be represented as a time-varying coefficient AR(1) model for $\log(h_t)$. Namely, after some simple algebra, by substituting the measurement equation in the volatility equation, the specification for the log-conditional variance can be rewritten as

$$\log(h_t) = \bar{\mu}_{t-1} + \bar{\pi}_{t-1} \log(h_{t-1}) + \bar{w}_{t-1} \tag{10}$$

where

$$\begin{aligned}
\bar{\mu}_t &= \omega + \gamma(\lambda_t \xi_H + (1-\lambda_t)\xi_L) \\
\bar{\pi}_t &= \beta + \gamma(\lambda_t \varphi_H + (1-\lambda_t)\varphi_L) \\
\bar{w}_t &= \gamma(\lambda_t w_{H,t} + (1-\lambda_t) w_{L,t})
\end{aligned}$$

and

$$\begin{aligned}
w_{j,t} &= u_{j,t} + \tau_j(z_t), \qquad j = H, L \\
u_{j,t} &= \log\left(x_t^{(j)}\right) - \xi_j - \varphi_j \log(h_t) - \tau_j(z_t), \qquad j = H, L.
\end{aligned}$$

Remark 3 (the SAF-RGARCH model). A more parsimonious version of the AF-RGARCH model can be obtained by collapsing the two measurement equations into one single equation in which the dependent variable is given by the weighted average $log(\tilde{x}_t)$. An appealing feature of this solution is that it allows to directly obtain insight on the statistical properties of the model based on the realized measure $\tilde{x}_t$, through the parameters of the dedicated measurement equation. The resulting specification is called the Single equation Adaptive Frequency Realized GARCH (SAF-RGARCH) model and it is defined by the following equations

$$\log(h_t) = \omega + \beta \log(h_{t-1}) + \gamma \log(\tilde{x}_{t-1}) \tag{11}$$

$$\log(\tilde{x}_t) = \lambda_t \log\left(x_t^{(H)}\right) + (1-\lambda_t) \log\left(x_t^{(L)}\right) \tag{12}$$

$$\tilde{u}_t = \log(\tilde{x}_t) - \xi - \varphi \log(h_t) - \tau(z_t). \tag{13}$$

The main characteristic of both the AF-RGARCH and SAF-RGARCH models is that they are able to incorporate information from different realized volatility measures based on intra-day returns observed at different frequencies, allowing for a

time-varying optimal bias/variance trade-off depending on the features of the two measures employed.

Remark 4 (market microstructure noise and attenuation bias). As discussed above, when prices are sampled at higher frequencies, microstructure problems become more pronounced. However, sampling at longer time horizons to obtain more reasonable estimates of volatility might not be an optimal solution. In this direction, the idea behind the Adaptive Frequency specifications follows both the results in Zhang et al. [37], by combining estimators obtained over two time scales to reduce market microstructure noise, and the findings in Gerlach et al. [22], by using time-varying weights to control the attenuation bias problem. Therefore, in addition to mitigating the market microstructure noise, accounting for the time-varying attenuation bias can potentially lead to better volatility and tail risk forecasts.

4 Principal Component Realized GARCH Models

It is worth remarking that the AF-RGARCH models allow mixing information from two different realized measures computed using intra-day returns sampled at different frequencies. However, it should be noted that the extension of these models to include an arbitrary number of frequencies is, for several reasons, not straightforward. On the other hand, an increasingly wide range of volatility estimators have been developed to control for market frictions. Thus, by crossing these two factors, estimation method and reference intra-day frequency, we obtain a potentially large number of volatility measures that can all be considered as noise-corrupted realizations of the same latent signal. In the presence of a large quantity of "signals" generated from the data under analysis, multivariate techniques such as PCA can be very useful for data reduction and pattern recognition.

In this section, we present extensions of the RGARCH model based on the use of PCA to reduce the dimensionality of a large set of realized volatility measures and achieve parsimony. The aim is to replace the single realized measure in the measurement equation of RGARCH with a new filtered "mixed" measure obtained by merging information from a possibly wide and representative set of different realized estimators relying on intra-day returns at different frequencies.

Along these lines, the Principal Component Realized GARCH (PC-RGARCH) is defined as

$$\log(h_t) = \omega + \beta \log(h_{t-1}) + \gamma \tilde{y}_{t-1} \tag{14}$$

$$\tilde{y}_t = \xi + \varphi \log(h_t) + \tau(z_t) + \tilde{u}_t \tag{15}$$

where $\tilde{y}_t$ is the first principal component resulting from a matrix of log-realized volatility measures generated by using different estimators and sampling frequencies.

When dealing with highly correlated data, as in the case of realized measures, PCA allows to obtain a lower-dimensional representation of the original dataset, without a substantial loss of information. As will be shown in the empirical analysis, since the fraction of variance explained by the first principal component is about 95%, the impact of succeeding components will not be considered in PC-RGARCH specifications. However, this does not preclude that if an additional component were to be needed, it could easily be introduced into the model by considering an additional measurement equation. For example, the same modelling approach proposed for AF-RGARCH or SAF-RGARCH could be adopted.

An interesting feature of PC-RGARCH is that, by preserving the same structure of the standard RGARCH, it allows the stochastic properties of the model to be easily derived. The goal of the feature selection specifications is to achieve the optimal compromise between bias and variability, in a data-driven fashion, thus eluding the problem of choosing the "best" estimator-frequency combination.

The PC-RGARCH, rather than relying on subjectively selected volatility measures, as with the RGARCH or Adaptive Frequency specifications, parsimoniously synthesizes the information provided by a panel of different realized measures, into a single or possibly a low number of components. As a result, by avoiding selecting a specific frequency or estimator for model fitting and forecasting, the model uncertainty associated with the RGARCH and (S)AF-RGARCH models is substantially reduced.

5 Estimation

In this section, we discuss the maximum likelihood estimation procedure of the Adaptive Frequency RGARCH and PC-RGARCH models.

Being characterized by a single measurement equation, the likelihood derivation of the SAF-RGARCH closely follows that of the standard Realized GARCH model. In our empirical application, differently from Hansen et al. [26], Student-t errors will be considered for the return equation.

The log-likelihood function of the SAF-RGARCH model is given by

$$\mathcal{L}(r, \tilde{x}; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log\, f(r_t, \tilde{x}_t | \mathcal{F}_{t-1})$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_h, \boldsymbol{\theta}_{\tilde{x}})'$ with $\boldsymbol{\theta}_h$ and $\boldsymbol{\theta}_{\tilde{x}}$ the vectors of parameters characterizing the volatility equation ($\boldsymbol{\theta}_h$) and the measurement equation ($\boldsymbol{\theta}_{\tilde{x}}$), respectively.

By standard probability theory results, this can be rewritten by factorizing the joint conditional density $f(r_t, \tilde{x}_t | \mathcal{F}_{t-1})$ as

$$f(r_t, \tilde{x}_t | \mathcal{F}_{t-1}) = f(r_t | \mathcal{F}_{t-1}) f(\tilde{x}_t | r_t; \mathcal{F}_{t-1}).$$

Finally, assuming a standardized Student-t distribution for $z_t \overset{iid}{\sim} t(0, 1, \nu)$ and a Gaussian distribution for $\tilde{u}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\tilde{u}}^2)$, the log-likelihood function of the SAF-RGARCH model is given by

$$\mathcal{L}(r, \tilde{x}; \boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{T} -\mathcal{A}(\nu) + \log(h_t) + (1+\nu)\,\log\left(1 + \frac{r_t^2}{h_t(\nu-2)}\right) \tag{16}$$

$$-\frac{1}{2}\sum_{t=1}^{T} \log(2\pi) + \log(\sigma_{\tilde{u}}^2) + \frac{\tilde{u}_t^2}{\sigma_{\tilde{u}}^2}, \tag{17}$$

where $\mathcal{A}(\nu) = \log\left[\Gamma\left(\frac{\nu+1}{2}\right)\right] - \log\left[\Gamma\left(\frac{\nu}{2}\right)\right] - \frac{1}{2}\,\log[\pi(\nu-2)]$. Equation (16) is the partial log-likelihood of the returns component, while Eq. (17) points out the contribution of realized measure component to the total log-likelihood value.

This log-likelihood structure can also be applied to PC-RGARCH specifications, replacing the selected realized measure $\log(x_t)$ by the synthetic realized measures given by the first principal component.

On the other hand, for the AF-RGARCHs, due to the presence of two measurement equations and, hence, two measurement errors, similarly to Hansen and Huang [25], we assume that $\mathbf{u}_t \overset{iid}{\sim} \mathcal{N}_2(\mathbf{0}, \Sigma)$, where $\mathcal{N}_2(\mathbf{0}, \Sigma)$ denotes a bivariate Normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ, with $\mathbf{u}_t = (u_{H,t}, u_{L,t})'$.

As a result, the log-likelihood function has the following structure

$$\mathcal{L}(r, u; \boldsymbol{\vartheta}) = -\frac{1}{2}\sum_{t=1}^{T} -\mathcal{A}(\nu) + \log(h_t) + (1+\nu)\,\log\left(1 + \frac{r_t^2}{h_t(\nu-2)}\right) \tag{18}$$

$$-\frac{1}{2}\sum_{t=1}^{T} K\log(2\pi) + \log\left(|\Sigma|\right) + \mathbf{u}_t'\,\Sigma^{-1}\,\mathbf{u}_t, \tag{19}$$

where K refers to the number of realized measures taken into account, that is $K = 2$ in our modelling approach.

It is worth noting that, for the proposed extensions of the RGARCH, the different specifications of the measurement equation make the models not directly comparable in terms of the overall maximized log-likelihood, but they are still comparable in terms of the partial log-likelihood of the returns component

$$\ell(r) = \sum_{t=1}^{T} \log\, f(r_t|\mathcal{F}_{t-1})$$

that, in general, can be used to compare the more sophisticated RGARCH models even with simple GARCH-type models based on end-of-day data.

6 Data Description

Our dataset consists of high-frequency prices of the Standard & Poor's 500 index, observed at frequencies ranging from 1-minute to 30-minute. In calculating daily open-to-close returns and realized measures, we focus on the period 03 January 2000–31 August 2016, limiting our attention to the official trading hours 9:30 am–4:00 pm. The data have been cleaned removing the last day of each year and some extreme outliers. In addition, for the computation of realized volatility and quarticity measures, the first and last observation of each trading day were excluded, as usual. As a result, the sample data consist of 4077 daily observations for each variable considered. Although several other estimators could be considered, in our empirical analysis, we build a panel of realized measures focusing on the *Realized Volatility* (RV) [7], *Realized Kernel* (RK) [10], *minRV* and *medRV* [8], *Realized BiPower Variation* (BPV) [13] and *Realized Outlyingness Weighted Covariation* (ROWCov) [14], computed at the 1, 3, 5, 10, 15 and 30-minute frequencies resulting in 36 different realized measures. In the selection process, the leading idea was to build a panel of realized measures that was representative of the main categories of realized volatility estimators typically used in financial applications: plain realized variance, micro-structure noise robust estimators and jump robust estimators.

Table 1 reports summary statistics of daily log-returns and of the considered log-transformed daily realized measures for the 5-minute sampling frequency. The daily log-returns show a standard deviation around 1% and are negatively skewed. Furthermore, the pronounced excess of kurtosis provides evidence for the non-Gaussianity of the r_t distribution, as expected. The realized measures, even after log-transformation, are still characterized by positive skewness and a moderate excess kurtosis.

Table 1 S&P500 index for the full sample period 03/01/2000–31/08/2016: summary statistics of log-returns and 5-min log-realized measures

	Mean	Std.dev	Median	Min	Max	Skewness	Kurtosis
r_t	0.000	0.010	0.000	−0.082	0.074	−0.150	7.334
$\log\left(RV_t^{(5)}\right)$	−9.971	1.047	−10.041	−12.940	−5.153	0.527	0.515
$\log\left(RK_t^{(5)}\right)$	−9.989	1.047	−10.058	−12.940	−5.156	0.510	0.486
$\log\left(medRV_t^{(5)}\right)$	−10.092	1.077	−10.177	−13.156	−5.263	0.529	0.495
$\log\left(minRV_t^{(5)}\right)$	−10.113	1.081	−10.198	−13.200	−5.472	0.540	0.523
$\log\left(BPV_t^{(5)}\right)$	−10.065	1.064	−10.146	−13.087	−5.324	0.536	0.516
$\log\left(ROWCov_t^{(5)}\right)$	−10.180	1.100	−10.264	−13.141	−5.392	0.467	0.372

Key to table. r_t: daily open-to-close log-returns; $RV_t^{(5)}$: daily 5-min log-Realized Volatility; $RK_t^{(5)}$: daily 5-min Realized Kernel; $medRV_t^{(5)}$: daily 5-min medRV; $minRV_t^{(5)}$: daily 5-min minRV; $BPV_t^{(5)}$: daily 5-min Realized BiPower Variation; $ROWCov_t^{(5)}$: daily 5-min log-Realized Outlyingness Weighted Covariation. Note that "Kurtosis" refers to excess of kurtosis

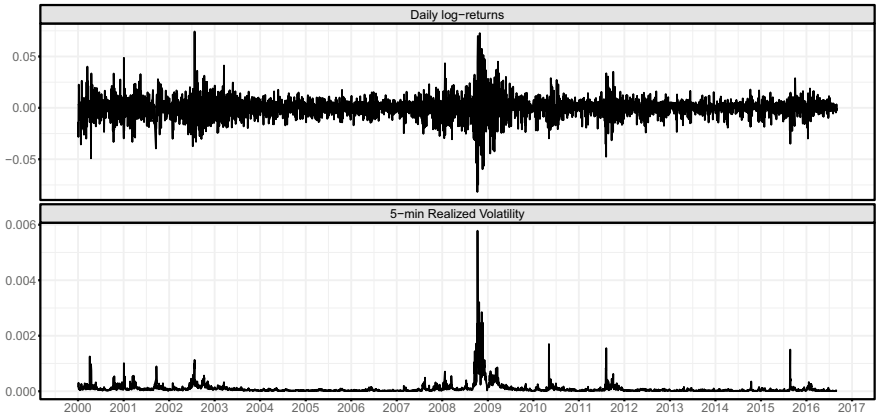


Fig. 1 Time series of daily log-returns and 5-min Realized Volatility Daily log-returns (top panel) and 5-minute Realized Volatility (bottom panel) of the S&P500 index for the full sample period 03/01/2000–31/08/2016

Table 2 Distribution of $log\left(R_t^{(H,L)}\right)$

	0%	5%	10%	25%	50%	75%	95%	95%	100%	Mean	Std.dev	Skew
$\log\left(R_t^{(1,5)}\right)$	−2.907	−1.014	−0.714	−0.272	0.186	0.639	1.107	1.491	6.416	0.705	2.485	3.745
$\log\left(R_t^{(3,5)}\right)$	−4.545	−0.603	−0.442	−0.186	0.121	0.440	0.762	1.004	4.307	0.296	0.376	3.470
$\log\left(R_t^{(5,10)}\right)$	−1.245	−0.635	−0.462	−0.164	0.173	0.526	0.898	1.126	3.652	0.378	0.259	2.333
$\log\left(R_t^{(5,15)}\right)$	−1.679	−0.693	−0.480	−0.095	0.334	0.814	1.301	1.613	7.195	0.836	3.045	4.133
$\log\left(R_t^{(5,30)}\right)$	−2.453	−0.772	−0.460	0.089	0.685	1.370	2.047	2.459	7.730	1.555	3.722	3.858

Summary of the distribution of the log-ratio $log\left(R_t^{(H,L)}\right) = \log\left(RQ_t^{(H)}/RQ_t^{(L)}\right)$, where $RQ_t^{(H)}$ and $RQ_t^{(L)}$ are realized quarticities based on a high and a low sampling frequency, respectively. The 5-minute RQ is taken as a fixed base

Figure 1 reports the time plots of the daily log-returns (top panel) and 5-min RV_t (bottom panel) of the S&P500 index for the full sample period 03 January 2000–31 August 2016. The analysis reveals several periods of high volatility, reflecting the impact of some extreme events: the bursting of the tech-bubble in the early 2000s; the financial crisis of 2007–2008, the crisis in Europe progressing from the banking system to the sovereign debt crisis, with the highest level of turbulence in late 2011; the stock market sell-off occurred between June 2015 and June 2016, along with the Chinese stock market turmoil, the uncertainty around FED interest rates, oil prices, Brexit and US presidential election.

Table 2 focuses on the features of $\log\left(R_t^{(H,L)}\right) = \log\left(RQ_t^{(H)}/RQ_t^{(L)}\right)$, which drives the evolution of the time-varying weights in the measurement equation of the Adaptive Frequency models. In our empirical application, this ratio was calculated by combining the usual 5-minute frequency [30] with other frequencies ranging from 1 to 30 min. It is worth noting that, since the frequency pairs involved are very close

Table 3 Summary of principal component analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	6.387	0.950	0.620	0.496	0.476	0.389	0.362	0.306	0.279	0.254
Proportion of Variance	0.940	0.021	0.009	0.006	0.005	0.003	0.003	0.002	0.002	0.001
Cumulative Proportion	0.940	0.961	0.970	0.976	0.981	0.984	0.987	0.989	0.991	0.993

The table shows the information about the first ten components of Principal Components Analysis on the data matrix of 36 log-realized measures: $\log\left(RV_t^{(s)}\right)$, $\log\left(RK_t^{(s)}\right)$, $\log\left(medRV_t^{(s)}\right)$, $\log\left(minRV_t^{(s)}\right)$, $\log\left(BPV_t^{(s)}\right)$ and $\log\left(ROWCov_t^{(s)}\right)$, with s =1, 3, 5, 10, 15, 30-minute

to each other, the log-ratios $R_t^{(5,10)}$ and $R_t^{(3,5)}$ show the lowest maximum values and standard deviations. On the other hand, the opposite occurs for $R_t^{(5,30)}$ which exhibits the highest maximum value and standard deviation. Furthermore, it can be easily seen that the greater the gap between the selected frequencies, the greater the percentage of $R_t^{(H,L)}$ observations that deviate from 1, as expected.

Finally, Table 3 provides information on the importance of the first ten components arising from the PCA based on the matrix given by 36 log-realized measures. The first principal component explains the 94% of the total variation in the data, while the second component explain approximately 2% of the overall variability. Thus, only the first principal component is chosen to represent our dataset. Consequently, in our empirical analysis, we will consider only the first latent component in the specification of PC-RGARCH model.

In addition, in order to assess the impact of the chosen frequency on tail risk forecasts, we consider two additional synthetic measures of volatility. Namely, we consider the first principal component computed using all available estimators at frequency greater than or equal to 5 min giving rise to the PC-RGARCH$^{(High)}$. Similarly, using all available estimators with frequency less than or equal to 5 min we obtain the PC-RGARCH$^{(Low)}$.

7 Tail-risk Forecasting

7.1 Forecasting Design

This section assesses the ability of the (S)AF-RGARCH and PC-RGARCH to generate accurate one-step-ahead forecasts of VaR and ES. The forecasting design is based on a rolling window scheme, with model parameters recursively estimated every day by ML. The out-of-sample period, ranging from 2 June 2008 to 31 August 2016, covers the 2008 credit crisis, the turmoil period related to the instability of the Euro area in late 2011 and the 2015/16 stock market sell-off, for a total of 2051 trading days. One-step-ahead VaR and ES forecasts are then generated for three different

risk levels, namely 1%, 2.5% and 5%. As a benchmark, the log-linear RGARCH model using the 5-minute RV is considered.

Formally, let $\mathcal{F}_t$ be the information available at time t and

$$F_{t,r}(x) = Pr(r_t \leq x|\mathcal{F}_{t-1})$$

be the Cumulative Distribution Function (CDF) of r_t conditional on $\mathcal{F}_{t-1}$. We assume that $F_{t,r}(\cdot)$ is strictly increasing and continuous on the real line $\mathbb{R}$. Under this assumption, the one-step-ahead α-level VaR at time t can be defined as

$$VaR_{t+1}(\alpha) = F^{-1}_{t+1,r}(\alpha) = h_{t+1}\, F^{-1}_{z}(\alpha) \equiv h_{t+1}\, z_\alpha, \qquad 0 < \alpha < 1,$$

where $F_z(\cdot)$ is the CDF of the returns innovations z_t. Under the same assumptions, the one-step-ahead α-level ES can be shown (see [2], among others) to be equal to the tail expectation of r_{t+1} conditional on a violation of $VaR_{t+1}(\alpha)$:

$$ES_{t+1}(\alpha) = E\left(r_{t+1}|\mathcal{F}_t, r_{t+1} \leq VaR_{t+1}(\alpha)\right) = h_{t+1}\, E(z_{t+1}|z_{t+1} \leq z_\alpha) \equiv h_{t+1}\, \mu_\alpha.$$

7.2 Evaluation of Tail-Risk Forecasts: Backtesting and Scoring Rules

First, the adequacy of VaR forecasts is assessed through backtesting procedures. Among the several tests that have been proposed to directly evaluate VaR quantile forecasts, here we focus on the Conditional Coverage (CC) test of Christoffersen [16] and on the Dynamic Quantile (DQ) test by Engle and Manganelli [18]. However, while these tests are useful for testing the adequacy of a given forecasting model, they do not allow to assess the statistical differences in forecasting performance between competing models. Accordingly, we rely on the Quantile Loss function [24] to rank models according to their ability to predict extreme losses:

$$QL_t(\alpha) = (\alpha - l_t)(r_t - VaR_t(\alpha)), \tag{20}$$

where l_t is a dummy variable such that $l_t = \mathcal{I}_{(r_t < VaR_t(\alpha))}$. The Quantile Loss ($QL$) function is well known to be a strictly consistent scoring rule for VaR prediction.

In addition, we complement our analysis by considering some "regulatory" loss functions that explicitly consider the magnitudes of violations when an exception occurs, accommodating regulators' concerns [1]. In particular, we consider the quadratic function of Lopez [31], given by

$$LPZ_t = \begin{cases} 1 + (r_t - VaR_t(\alpha))^2 & \text{for } r_t < VaR_t(\alpha) \\ 0 & \text{for } r_t \geq VaR_t(\alpha). \end{cases} \tag{21}$$

Caporin [15] extends this idea considering different weighting functions for the magnitude of the violations, leading to the three following alternative loss functions

$$RC1_t = \begin{cases} \left|1 - \left|\frac{r_t}{VaR_t(\alpha)}\right|\right| & \text{for } r_t < VaR_t(\alpha) \\ 0 & \text{for } r_t \geq VaR_t(\alpha) \end{cases} \tag{22}$$

$$RC2_t = \begin{cases} \frac{(|r_t| - |VaR_t(\alpha)|)^2}{VaR_t(\alpha)} & \text{for } r_t < VaR_t(\alpha) \\ 0 & \text{for } r_t \geq VaR_t(\alpha) \end{cases} \tag{23}$$

$$RC3_t = \begin{cases} |r_t - VaR_t(\alpha)| & \text{for } r_t < VaR_t(\alpha) \\ 0 & \text{for } r_t \geq VaR_t(\alpha). \end{cases} \tag{24}$$

As opposed to VaR, ES lacks the mathematical property called *elicitability* [20, 21, 23, 36], since there is no loss function available for which the ES is the solution to minimizing the expected loss. However, Fissler and Ziegel [20] show that VaR and ES are jointly elicitable with respect to the following class of loss functions

$$FZ_t = (l_t - \alpha)\left(G_1(v_t) - G_1(r_t) + \frac{1}{\alpha}G_2(e_t)v_t\right) - G_2(e_t)\left(\frac{1}{\alpha}l_t r_t - e_t\right) - \mathcal{G}_2(e_t), \tag{25}$$

where $G_1(\cdot)$ is weakly increasing, $G_2(\cdot)$ is strictly increasing and strictly positive and $\mathcal{G}_2'(\cdot) = G_2(\cdot)$, with v_t and e_t, the VaR and ES, respectively. Several strictly consistent scoring rules for the couple (VaR, ES) can be derived as special cases of the family of functions in (25). In this paper, assuming VaR and ES to be strictly negative, with $ES_t(\alpha) \leq VaR_t(\alpha) < 0$, we consider the zero-degree homogeneous loss function as defined in Patton et al. [33], obtained from (25) by setting $G_1(x) = 0$ and $G_2(x) = -1/x$

$$FZ_t^{(0)} = \frac{1}{\alpha ES_t(\alpha)} l_t \left(r_t - VaR_t(\alpha)\right) + \frac{VaR_t(\alpha)}{ES_t(\alpha)} + \log(-ES_t(\alpha)) - 1. \tag{26}$$

As for the other loss functions, models characterized by lower average values of $FZ_t^{(0)}$ in the out-of-sample forecast evaluation period are preferred.

The significance of differences in forecasting performance across different models has been tested by the Model Confidence Set (MCS) approach of Hansen et al. [28] considering the confidence levels of 75% and 90% according to the SQ (Semi-Quadratic) statistic.[1] The procedure has been implemented using a block-bootstrap procedure based on 5000 resamples [32].

[1] We do not show the R (Range) statistic, as it gives practically the same results as for the SQ statistic.

Table 4 Value-at-Risk backtesting. VR shows the Violation Rate as proportion of returns smaller than VaR during the forecast period (2051 days) at the risk levels of 1%, 2.5% and 5%. CC and DQ report the p-values for the Conditional Coverage test and Dynamic Quantile test, respectively. In **bold** models showing the VR closest to the nominal level α

	$\alpha = 0.01$			$\alpha = 0.025$			$\alpha = 0.05$		
	VR	CC	DQ	VR	CC	DQ	VR	CC	DQ
RG$^{(5)}$	0.0137	0.1964	0.1605	0.0317	0.1231	0.5219	0.0556	0.4158	0.7743
SAF-RG$^{(5,1)}$	0.0146	0.0920	0.0923	**0.0288**	0.4739	0.2294	**0.0502**	0.2693	0.1452
SAF-RG$^{(5,3)}$	0.0137	0.1964	0.1653	0.0302	0.2621	0.2683	0.0561	0.2790	0.4009
SAF-RG$^{(5,10)}$	0.0141	0.1368	0.1126	0.0317	0.1231	0.5222	0.0551	0.4473	0.5646
SAF-RG$^{(5,15)}$	0.0141	0.1368	0.1146	0.0322	0.0923	0.4421	0.0561	0.3820	0.6739
SAF-RG$^{(5,30)}$	0.0137	0.1964	0.1604	0.0322	0.0923	0.4415	0.0561	0.3820	0.6745
AF-RG$^{(5,1)}$	0.0137	0.1964	0.1692	**0.0288**	0.4739	0.3150	0.0561	0.2790	0.4204
AF-RG$^{(5,3)}$	**0.0132**	0.2715	0.1846	0.0312	0.1612	0.5083	0.0570	0.2366	0.3719
AF-RG$^{(5,10)}$	0.0141	0.1368	0.1157	0.0317	0.1231	0.5213	0.0556	0.4158	0.6979
AF-RG$^{(5,15)}$	0.0141	0.1368	0.1098	0.0322	0.0923	0.4409	0.0556	0.4158	0.7638
AF-RG$^{(5,30)}$	0.0137	0.1964	0.1669	0.0307	0.2074	0.6038	0.0561	0.3820	0.7972
PC-RG	0.0141	0.1368	0.1128	0.0317	0.1231	0.4266	0.0585	0.1196	0.2226
PC-RG$^{(High)}$	**0.0132**	0.2715	0.1826	0.0293	0.3963	0.3301	0.0551	0.4473	0.6093
PC-RG$^{(Low)}$	0.0146	0.0920	0.1146	0.0332	0.0494	0.2399	0.0580	0.1327	0.1291

7.3 Empirical Results

Table 4 reports the results of the backtesting procedure applied to VaR forecasts. For each risk level, the table shows the empirical Violation Rate (VR), that is the proportion of returns smaller than VaR within the forecast period, and the p-values of the Conditional Coverage (CC) and Dynamic Quantile (DQ) tests. A correctly specified model should present an empirical VR close to the specified VaR level. In this direction, Table 4 shows VR values quite close to the nominal risk level, with the DQ and CC test statistics always lying in the acceptance region at the usual 5% level (except for PC-RG$^{(Low)}$ at the 0.025 risk level using the CC test).

Table 5 reports, for the three considered risk levels, the average values of the QL loss function, together with the associated MCS p-values. The results show that the QL is always minimized by the PC-RGARCH class, with the PC-RGARCH$^{(Low)}$ the only model that always enters the 75% MCS, regardless of the risk level considered. On the other hand, the predictive performance of the Adaptive Frequency models depends on the chosen frequency combination. In particular, it turns out that for extreme risk levels, specifications that mix low-frequency measures tend to be preferred. The opposite occurs when moving toward more moderate levels of risk, such as 5%, with higher frequency combinations returning lower values of the loss function. In addition, as shown by the MCS results, the most parsimonious SAF-RGARCH models provide performance in line with that of AF-RGARCH. Finally,

Table 5 Quantile Loss function at different risk levels. For each $VaR_t(\alpha)$ series, with $\alpha = 0.01, 0.025, 0.05$, the table shows the average values of Quantile Loss (QL) function and the MCS p-value. In **bold** models minimizing the loss. In box models $\in$ 90% MCS and in box models $\in$ 75% MCS

	$\alpha = 0.01$		$\alpha = 0.025$		$\alpha = 0.05$	
	QL	p-value	QL	p-value	QL	p-value
RG	0.5377	0.1182	1.1607	0.0710	2.0015	0.0834
SAF-RG$^{(5,1)}$	0.5568	0.0540	1.1820	0.0662	1.9903	0.3696
SAF-RG$^{(5,3)}$	0.5414	0.0850	1.1528	0.1460	1.9903	0.3696
SAF-RG$^{(5,10)}$	0.5388	0.1080	1.1612	0.0710	2.0004	0.0834
SAF-RG$^{(5,15)}$	0.5375	0.1182	1.1608	0.0710	2.0008	0.0834
SAF-RG$^{(5,30)}$	0.5380	0.1182	1.1597	0.0710	2.0009	0.0834
AF-RG$^{(5,1)}$	0.5498	0.0432	1.1678	0.0710	1.9861	0.3696
AF-RG$^{(5,3)}$	0.5384	0.1132	1.1498	0.2130	1.9929	0.3286
AF-RG$^{(5,10)}$	0.5372	0.1182	1.1601	0.0710	2.0021	0.0834
AF-RG$^{(5,15)}$	0.5381	0.1182	1.1593	0.0738	1.9986	0.1220
AF-RG$^{(5,30)}$	0.5354	0.1182	1.1603	0.0902	2.0002	0.0874
PC-RG	0.5313	0.1182	**1.1391**	1.0000	1.9831	0.5836
PC-RG$^{(High)}$	0.5369	0.1182	1.1495	0.2980	**1.9734**	1.0000
PC-RG$^{(Low)}$	**0.5278**	1.0000	1.1450	0.5106	1.9967	0.3696

the benchmark RGARCH model only enters the less restrictive 90% MCS at the 1% risk level.

The analysis based on the regulatory loss functions gives a slightly different picture. It is here worth noting that, in general, the aim of regulatory loss functions, such as those considered in this paper, is to rank models according to their ability to match regulators' concerns rather than according to their forecasting accuracy. They, in fact, exclusively focus on measuring the average magnitude of violations. Under this respect, the results in Table 6 show that the best results are achieved when using the SAF-RGARCH$^{(1,5)}$ model, minimizing the loss in eight cases out of twelve. In more detail, at the 5% risk level, all four regulatory losses considered are minimized by the SAF-RGARCH$^{(1,5)}$, while, at the 2.5% risk level, this occurs in three out of four cases, with the PC-RGARCH minimizing the $RC1$ loss. At the 1% risk level, we find that SAF-RGARCH$^{(1,5)}$ takes the lowest values for $RC3$, while $RC1$ and $RC2$ are minimized by PC-RGARCH$^{(Low)}$ and PC-RGARCH$^{(High)}$, respectively. For LPZ, the minimum loss value is obtained by AF-RGARCH$^{(3,5)}$ and PC-RGARCH$^{(High)}$.

Finally, we focus on the ability of the compared models to generate forecast joint predictions of the pair (VaR, ES). In this respect, Table 7 reports the $FZ_t^{(0)}$ losses averaged over the forecasting period for each specification along with the MCS p-values. Models with lower average losses are preferred, while the larger the p-value, the higher a model is ranked from the MCS procedure. Similarly to what we found for the QL in Table 5, the PC-RGARCH specifications provide lower

Table 6 VaR regulatory Loss Functions at different risk levels. For each $VaR_t(\alpha)$ series, with $\alpha = 0.01, 0.025, 0.05$, the table shows the loss values for LPZ^* ($*$: average loss $\times 100$), $RC1^*$ ($*$: average loss $\times 100$), $RC2$ and $RC3$, namely the Lopez and the three Caporin loss functions. In **bold** models minimizing the losses

	$\alpha = 0.01$				$\alpha = 0.025$				$\alpha = 0.05$			
	LPZ^*	$RC1^*$	$RC2$	$RC3$	LPZ^*	$RC1^*$	$RC2$	$RC3$	LPZ^*	$RC1^*$	$RC2$	$RC3$
RG$^{(5)}$	1.3652	0.3071	0.0307	0.0964	3.1693	0.9340	0.1165	0.2635	5.5585	2.1002	0.3156	0.5318
SAF-RG$^{(5,1)}$	1.4627	0.2991	0.0286	**0.0879**	**2.8767**	0.8784	**0.1076**	**0.2303**	**5.0221**	**1.9424**	**0.2789**	**0.4334**
SAF-RG$^{(5,3)}$	1.3652	0.3126	0.0317	0.0960	3.0230	0.9276	0.1161	0.2473	5.6072	2.0845	0.3057	0.5068
SAF-RG$^{(5,10)}$	1.4140	0.3090	0.0311	0.0965	3.1693	0.9344	0.1172	0.2622	5.5097	2.0981	0.3159	0.5279
SAF-RG$^{(5,15)}$	1.4140	0.3061	0.0302	0.0959	3.2180	0.9314	0.1158	0.2630	5.6072	2.0940	0.3143	0.5301
SAF-RG$^{(5,30)}$	1.3652	0.3106	0.0306	0.0967	3.2180	0.9362	0.1163	0.2623	5.6072	2.1056	0.3149	0.5305
AF-RG$^{(5,1)}$	1.3652	0.3013	0.0288	0.0916	**2.8767**	0.8893	0.1099	0.2368	5.6072	1.9832	0.2857	0.4615
AF-RG$^{(5,3)}$	**1.3165**	0.3086	0.0311	0.0956	3.1205	0.9202	0.1149	0.2495	5.7047	2.0856	0.3077	0.5180
AF-RG$^{(5,10)}$	1.4140	0.3054	0.0307	0.0963	3.1693	0.9286	0.1155	0.2632	5.5585	2.0897	0.3134	0.5325
AF-RG$^{(5,15)}$	1.4140	0.3092	0.0325	0.0978	3.2180	0.9204	0.1173	0.2639	5.5585	2.0740	0.3149	0.5315
AF-RG$^{(5,30)}$	1.3652	0.2976	0.0300	0.0950	3.0718	0.9169	0.1153	0.2649	5.6073	2.0780	0.3147	0.5334
PC-RG	1.4140	0.2825	0.0290	0.0902	3.1693	**0.8686**	0.1087	0.2422	5.8510	2.0183	0.2965	0.5139
PC-RG$^{(High)}$	**1.3165**	0.2928	**0.0282**	0.0893	2.9255	0.8936	0.1082	0.2393	5.5097	2.0047	0.2875	0.4823
PC-RG$^{(Low)}$	1.4627	**0.2783**	0.0296	0.0908	3.3155	0.8743	0.1118	0.2567	5.8023	2.0592	0.3093	0.5418

Table 7 $FZ^{(0)}$ loss function at different risk levels. For each $(VaR_t(\alpha), ES_t(\alpha))$ series, with $\alpha = 0.01, 0.025, 0.05$, the table shows the average values of $FZ^{(0)}$ loss function and the MCS p-value. In **bold** models minimizing the loss. In box models $\in 90\%$ MCS and in box models $\in 75\%$ MCS

	$\alpha = 0.01$		$\alpha = 0.025$		$\alpha = 0.05$	
	$FZ^{(0)}$	p-value	$FZ^{(0)}$	p-value	$FZ^{(0)}$	p-value
RG$^{(5)}$	-3.7302	0.0638	-3.8920	0.0226	-4.0639	0.0790
SAF-RG$^{(5,1)}$	-3.7111	0.0540	-3.8852	0.0244	-4.0643	0.2366
SAF-RG$^{(5,3)}$	-3.7254	0.0638	-3.8941	0.0244	-4.0663	0.1524
SAF-RG$^{(5,10)}$	-3.7280	0.0568	-3.8913	0.0198	-4.0637	0.0728
SAF-RG$^{(5,15)}$	-3.7316	0.0746	-3.8933	0.0244	-4.0652	0.1418
SAF-RG$^{(5,30)}$	-3.7284	0.0638	-3.8922	0.0244	-4.0637	0.0970
AF-RG$^{(5,1)}$	-3.7178	0.0568	-3.8895	0.0244	-4.0652	0.1524
AF-RG$^{(5,3)}$	-3.7296	0.0638	-3.8974	0.0510	-4.0672	0.1524
AF-RG$^{(5,10)}$	-3.7311	0.0638	-3.8928	0.0244	-4.0643	0.1098
AF-RG$^{(5,15)}$	-3.7278	0.0638	-3.8959	0.0244	-4.0675	0.1564
AF-RG$^{(5,30)}$	-3.7389	0.1092	-3.8981	0.0294	-4.0678	0.1524
PC-RG	-3.7528	0.2070	-3.9144	0.7446	**-4.0776**	1.0000
PC-RG$^{(High)}$	-3.7382	0.1512	-3.9006	0.1220	-4.0739	0.4480
PC-RG$^{(Low)}$	**-3.7593**	1.0000	**-3.9158**	1.0000	-4.0748	0.4480

losses than the competitors and are the only models always included in the MCS. In particular, for $\alpha = 0.01$, the only model entering the MCS together with the PC-RGARCH models is the AF-RGARCH$^{(5,30)}$. For the 2.5% risk level, no model other than PC-RGARCH, PC-RGARCH$^{(High)}$ and PC-RGARCH$^{(Low)}$ enters the MCS at the considered confidence levels of 75 and 90%. Finally, even for $\alpha = 0.05$, the only three models that enter the 75% MCS are the RGARCH that use synthetic realized volatility measures, while some other SAF-RGARCH and AF-RGARCH models enter the 90% MCS. In addition, the results for joint predictions of the pair (VaR, ES) confirm that moving from α =0.01 to α =0.05, models based on higher frequency measures provide a better forecasting performance. Again, the standard RGARCH is never included in the MCS.

8 Conclusion

In this paper, we have investigated the implications for tail risk forecasting of combining realized volatility measures characterized by different sensitivity to jumps and market microstructure noise. To this end, both different categories of estimators and multiple sampling frequencies for intra-day returns were considered. In this framework, two different extensions of the standard RGARCH model have been proposed. The first modelling solution allows to exploit the time-varying information from two realized volatility measures based on intra-day returns observed at different frequencies and has the role of correcting for upward and downward biases that could affect the dynamics of the measures involved. This class of models includes the Adaptive Frequency specifications, i.e., the AF-RGARCH and the more parsimonious SAF-RGARCH obtained by replacing the two measurement equations with a single equation. The second solution leads to features selection based PC-RGARCH models that can be represented as RGARCH models that rely on synthetic realized measures obtained through the application of the PCA. The use of dimension reduction techniques allows to reduce the dimensionality of a large panel of realized measurements without any substantial loss of information and, at the same time, allows to achieve parsimony and to mitigate the modelling uncertainty in selecting the "best" estimator-frequency combination. The out-of-sample forecasting results provide evidence that combining several realized measures, possibly computed using intra-day returns observed at different frequencies, leads to significant accuracy gains in forecasting tail risk, for both VaR and ES. In terms of forecast accuracy, models based on feature selection methods outperform both the standard RGARCH and Adaptive Frequency specifications. This is due to the greater flexibility of the PC-RGARCH models to parsimoniously merge information from several realized estimators computed over a possibly large set of frequencies. On the other hand, Adaptive Frequency models show a relatively good performance when the focus is on the magnitude of VaR violations. As a direction for future research, it would be interesting to consider a wider set of realized estimators, such as the Two Time Scale and the Realized Range, but also different sampling frequencies to be used both as a benchmark in the

Realized GARCH model and as a vehicle for additional information in the Adaptive Frequency and PCA models.

References

1. Abad, P., Muela, S.B., Martín, C.L.: The role of the loss function in value-at-risk comparisons. J. Risk Model Validation **9**(1), 1–19 (2015)
2. Acerbi, C., Tasche, D.: On the coherence of expected shortfall. J. Bank. Finance **26**(7), 1487–1503 (2002)
3. Aït-Sahalia, Y., Mykland, P.A., Zhang, L.: How often to sample a continuous-time process in the presence of market microstructure noise. Rev. Financ. Stud. **18**(2), 351–416 (2005)
4. Aït-Sahalia, Y., Yu, J.: High frequency market microstructure noise estimates and liquidity measures. Ann. Appl. Stat. **3**(1), 422–457 (2009)
5. Alexander, C.: Market Models: A Guide to Financial Data Analysis. Wiley (2001)
6. Andersen, T.G., Bollerslev, T.: Answering the skeptics: yes, standard volatility models do provide accurate forecasts. Int. Econ. Rev. 885–905 (1998)
7. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: Modeling and forecasting realized volatility. Econometrica **71**(2), 579–625 (2003)
8. Andersen, T.G., Dobrev, D., Schaumburg, E.: Jump-robust volatility estimation using nearest neighbor truncation. J. Econometrics **169**(1), 75–93 (2012)
9. Bandi, F.M., Russell, J.R.: Microstructure noise, realized variance, and optimal sampling. Rev. Econ. Stud. **75**(2), 339–369 (2008)
10. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica **76**(6), 1481–1536 (2008)
11. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Realized kernels in practice: trades and quotes. Econometrics J. **12**(3), (2009)
12. Barndorff-Nielsen, O.E., Shephard, N.: Econometric analysis of realized covariation: high frequency based covariance, regression, and correlation in financial economics. Econometrica **72**(3), 885–925 (2004)
13. Barndorff-Nielsen, O.E., Shephard, N.: Power and bipower variation with stochastic volatility and jumps. J. Financ. Econometrics **2**(1), 1–37 (2004)
14. Boudt, K., Croux, C., Laurent, S.: Outlyingness weighted covariation. J. Financ. Econometrics **9**(4), 657–684 (2011)
15. Caporin, M.: Evaluating value-at-risk measures in the presence of long memory conditional volatility. J. Risk **10**(3), 79 (2008)
16. Christoffersen, P.F.: Evaluating interval forecasts. Int. Econ. Rev. 841–862 (1998)
17. Contino, C., Gerlach, R.H.: Bayesian tail-risk forecasting using realized Garch. Appl. Stoch. Models Bus. Ind. **33**(2), 213–236 (2017)
18. Engle, R.F., Manganelli, S.: Caviar: conditional autoregressive value at risk by regression quantiles. J. Bus. Econ. Stat. **22**(4), 367–381 (2004)
19. Fan, J., Wang, M., Yao, Q.: Modelling multivariate volatilities via conditionally uncorrelated components. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **70**(4), 679–702 (2008)
20. Fissler, T., Ziegel, J.F.: Higher order elicitability and Osband's principle. Ann. Stat. **44**(4), 1680–1707 (2016)
21. Fissler, T., Ziegel, J.F., Gneiting, T.: Expected shortfall is jointly elicitable with value at risk-implications for backtesting. Risk Mag. 58–61 (2016)
22. Gerlach, R., Naimoli, A., Storti, G.: Time-varying parameters realized garch models for tracking attenuation bias in volatility dynamics. Quant. Finance **20**(11), 1849–1878 (2020)
23. Gneiting, T.: Making and evaluating point forecasts. J. Am. Stat. Assoc. **106**(494), 746–762 (2011)

24. González-Rivera, G., Lee, T.-H., Mishra, S.: Forecasting volatility: a reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. Int. J. Forecast. **20**(4), 629–645 (2004)
25. Hansen, P.R., Huang, Z.: Exponential garch modeling with realized measures of volatility. J. Bus. Econ. Stat. **34**(2), 269–287 (2016)
26. Hansen, P.R., Huang, Z., Shek, H.H.: Realized garch: a joint model for returns and realized measures of volatility. J. Appl. Econometrics **27**(6), 877–906 (2012)
27. Hansen, P.R., Lunde, A.: Realized variance and market microstructure noise. J. Bus. Econ. Stat. **24**(2), 127–161 (2006)
28. Hansen, P.R., Lunde, A., Nason, J.M.: The model confidence set. Econometrica **79**(2), 453–497 (2011)
29. Li, Y.-N., Zhang, Y., Zhang, C.: Statistical inference for measurement equation selection in the log-realgarch model. Econometric Theory **35**(5), 943–977 (2019)
30. Liu, L.Y., Patton, A.J., Sheppard, K.: Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. J. Econometrics **187**(1), 293–311 (2015)
31. Lopez, J.: Methods for evaluating value-at-risk estimates. Econ. Rev. 3–17 (1999)
32. Patton, A., Politis, D.N., White, H.: Correction to "automatic block-length selection for the dependent bootstrap" by d. politis and h. white. Econometric Rev. **28**(4), 372–375 (2009)
33. Patton, A.J., Ziegel, J.F., Chen, R.: Dynamic semiparametric models for expected shortfall (and value-at-risk). J. Econometrics **211**(2), 388–413 (2019)
34. Van der Weide, R.: Go-garch: a multivariate generalized orthogonal garch model. J. Appl. Econometrics **17**(5), 549–564 (2002)
35. Watanabe, T.: Quantile forecasts of financial returns using realized garch models. Japan. Econ. Rev. **63**(1), 68–80 (2012)
36. Weber, S.: Distribution-invariant risk measures, information, and dynamic consistency. Math. Finance: Int. J. Math. Stat. Financ. Econ. **16**(2), 419–441 (2006)
37. Zhang, L., Mykland, P.A., Aït-Sahalia, Y.: A tale of two time scales: determining integrated volatility with noisy high-frequency data. J. Am. Stat. Assoc. **100**(472), 1394–1411 (2005)

Multiversal Methods in Observational Studies: The Case of COVID-19

Venera Tomaselli, Giulio Giacomo Cantone, and Vincenzo Miracula

Abstract In the present study, 13 covariates have been selected as potentially associated with 3 metrics of the spread of COVID-19 in 20 European countries. Robustness of the linear correlations between 10 of the 13 covariates as main regressors and the 3 COVID-19 metrics as dependent variables have been tested through a methodology for sensitivity analysis that falls under the name of "Multiverse". Under this methodology, thousands of alternative estimates are generated by a single hypothesis of regression. The capacity of identification of a robust causal claim for the 10 variables has been measured through 3 indicators over a Janus Confusion Matrix, which is a confusion matrix that assumes the likelihood to observe a True claim as the ratio between the absolute difference of estimates with a different sign and the total of estimates. This methodology provides the opportunity to evaluate the outcomes of a shift from the common level of significance $\alpha = .05$ to the alternative $\alpha = .005$. According to the results of the study, in the dataset the benefits of the shifts come at a very high cost in terms of false negatives.

Keywords Multiverse analysis · Model mis-specification · *p*-hacking, significance level, COVID-19

V. Tomaselli (✉)
Department of Political and Social Sciences, University of Catania, 8, Vittorio Emanuele II, 95131 Catania, Italy
e-mail: venera.tomaselli@unict.it

G. G. Cantone · V. Miracula
Department of Physics and Astronomy "E. Majorana", University of Catania, 64, S. Sofia, 95123 Catania, Italy
e-mail: giulio.cantone@phd.unict.it

V. Miracula
e-mail: vincenzo.miracula@phd.unict.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_22

1 Introduction

In hypothesis testing, the probability of randomly drawing, in the set of theoretical circumstances described under the null hypothesis, a value as extreme as the value empirically observed, is referred as the p-value. Historically, a p-value smaller than the level of $\alpha = 0.05$ signalled a *statistically significant* the result of the test, which means that the presented evidence is not compatible with the null hypothesis. In this case the researcher has a justification to reject the null hypothesis [87].

Over time, the status of p-values in scientific research reached a situation of paradox: null hypothesis statistical testing (usually with a significance level $\alpha < 0.05$) is the most taught and the most used method for scientific inference, and at the same time, by its broad use, it started to be regarded as the main culprit for the lack of reliability in science [30, 55, 67, 82]. Skepticism and self-criticism towards standard scientific practices raised after the mediatic emergence of the so called 'replicability crisis' that invested psychological research [2, 10, 21, 52, 58, 85], health sciences [4, 34, 36], and is a concern for other fields [2, 6, 20, 61, 80, 91].

While there are methodological reasons why a null hypothesis statistical testing could lead into a lack of not replicated scientific claims, the focus is on the noxious practice of 'p-hacking' [33, 54, 77]. It consists in collecting a large amount of model specifications of a scientific claim in order to randomly see a significant result popping up by chance, then not reporting the exact number of attempts before reaching a $p < \alpha$.

Many scholars [35, 38, 48, 83, 88] saw an opportunity for a reflection about the possibilities for an advancement beyond the rule the *status quo* of hypothesis testing with $\alpha < 0.05$. A widely discussed proposal is to lower conventional α into 0.005 [5, 37, 42]. This proposal would not require to upgrade introductory classes in Statistics towards Bayesian or other methodologies. However, the raise in the false negative rate could prevent researcher to pursue highly innovative hypotheses. Also, to lower the α is virtually useless against the most extensive forms of p-hacking, e.g., on massive datasets with hundreds of variables (Big Data).

The issue of p-hacking inspired a different approach that falls under the name of multiversal methodology or 'multiverse-style methods' [18]. This methodology aims at observing a multivariate population (usually consisting of p-values and estimates $\hat{\beta}_x$) throughout mapping every single reasonable model specification of a causal relationship between a dependent variable y and a set of regressors (x, Z). This population is referred as 'the multiverse' of a study.

The multiverse tells something about the sensitivity of specific hypotheses $x \rightarrow y$ to their alternative specifications or *vibrations* [65]: multiversal statistics are informative regarding the practical possibility to p-hack the hypothesis or to incur false positives due mis-specification.

In the present study, the multiversal methods are discussed to estimate coefficients of multiple specifications of regressive models in Sect. 2. In Sect. 3, population, mobility, pollution, and public health variables in European countries and COVID-

19 data are selected for the analysis. In Sect. 4, the multiverse models are employed to data analysis. Lastly, in Sect. 5, multiverse models provide evidences about the COVID-19 pandemic spread and its effects on the health-care systems.

2 Theory: Multiversal Methods

All the multiversal methods are based on the estimation of coefficients of a large number of regressive models organised as a 'family'. The technical premises are the same of Extreme Bound Analysis (EBA) [29, 46, 71] but research questions are broader. The theoretical connection between the number of different attempts to 'make a model work' and the robustness of its scientific claim was made explicit in [77] and [28], which popularised the concept of 'Researcher's degree of freedom' with the metaphor of the "Garden of Forking Paths", a literary invention of novelist Borges.

A multiverse of specifications can be analysed through plotting:

- The Multiverse grid [32, 81] is a multidimensional array with all the specifications represented by their p-values, clustered in the grid space by the divergences in the 'Garden of Forking Paths'. Significant p-values are highlighted. This tool is impractical for a high number of specifications and is usually uninformative on the estimate.
- The Vibration-of-Effect (VoE) plot [65], a cartesian representation where
  - in the x-axis are represented the estimates of a standardised regressor x in the specifications of the multiverse,
  - in the y-axis are represented the logarithms of $log_{10}(p)$ associated to the null-hypothesis of coefficient equal to 0, multiplied $per\ -1$.

  VoE is used to display the sensitivity of a causal relationship between regressor and dependent as the p-value decreases. At the same time, it allows to detect the so-called "Janus effect", i.e., the sign discordance between estimates in the same family of specifications.
- The p-curve [7, 81] is a representation of the probability density associated to multiversal p-values: the more the density in p-curve is right-skewed towards lower p-values, the more the regressor is validated.
- The Specification Curve [78], which allows to compare different families of specifications (i.e., aggregations of micro-variants of a single causal hypothesis) and to associate these families both to an interval of p-values and to an interval of estimates.

Worth to mention in the family of multiversal methods is the Computational Framework for Multimodel Analysis [56, 92], which is an alternative to Machine Learning for model selection in a multivariate context.

An interesting application of the multiverse analysis is in [72]. In that paper the three authors had *divergent* results regarding the determinants of political behaviour of U.S. representatives. Instead of defending own theses to the bitter, the three shared their methodological designs (e.g., how to collect data), chose to collaborate to investigate the robustness of the divergent hypotheses through a multiverse of specifications (in the jargon of [65], they 'vibrated' them), and finally converged into a unique set of scientific claims.

Other applications of multiverse analysis in empirical research are in [16, 50, 59, 70].

A software to perform Specification Curve Analysis and, more in general, to generate a multiverse from a dataset is *specr* [51][1].

However, a properly unified multiversal methodology is still in development. The most theoretical contribution to the topic is by Del Giudice and Gangestad [18]. While the authors highlight both promising features and pitfalls of "The Multiverse", their main concern is with the phase of analytical choice in order to differentiate (*vibrate*) the specification of a hypothesis into families of model specifications. The argument provided by Del Giudice and Gangestad follows the more known scientific controversy regarding the introduction of a collider variable as a control in a regression model [48].

In the context of multiverse analysis, the controversy could be simplified to only a question: what is a reasonable vibration for a hypothesis?

Del Giudice and Gangestad [18] discuss about the covariates' selection as a basic issue in the literature. According to Simonsohn et al. [79] the covariates are linked to the chance to provide different answers to different research hypotheses. On the contrary, Patel et al. [65] demonstrate that the VoE emerges only with robustness analyses involving selected and alternative covariates. These controversial claims show that the lacking of agreement about clear and accurate guidelines does not allow to increase the potential of multiverse methods in data analysis.

2.1 What Is a Specification?

A model specification of the causal relationship $x \rightarrow y$:

$$y = f(x, Z) \tag{1}$$

is a member of the family of regressive models formalised as:

$$\overleftrightarrow{y}_{k_y} \sim \overleftrightarrow{F}_{k_F}(\overleftrightarrow{x}_{k_x}, \overleftrightarrow{Z}_{k_Z}) + \epsilon \tag{2}$$

where $\leftrightarrow$ indicates an operationalization, i.e., a decision to represent a theoretical concept (e.g., a statistical population) through a full identified object (e.g., an empiric

[1] Suggested tutorial: https://dcosme.github.io/specification-curves/SCA_tutorial_inferential.

sample). In Eq. 1, Z is a set of control covariates that the researchers necessary deem for the correct causal inference. In Eq. 2, F represents the set of equivalent functions that link the joint predictors x and $z \in Z$ to the outcomes y. k are indices for single operationalizations of the respective constructs [78].

An interesting propriety of the operationalization is that it does not only involve decisions about what to measure but also about *how* to measure it. Indeed, a core element of the methodology in Simonsohn, Simmons, and Nelson [78] is that they stress the importance of recoding the same observations through different scales of measurements.

In the literature, five elements of a specification are often reported:

1. The **Subsetting** of observations: here the decision regards mostly the inclusion of outliers or other peculiar clusters of observations. However, as a general rule, finding a reasonable criterion to split the dataset into subsets should help to assert the sensitivity of the relationship [3].
2. The **Regressors** (x, Z): this operationalization can be split into two different decisions:

    a. one regards the controversy about full inclusion of all the $n!$ combinations of the n covariates as regressors x (and/or controls) or to make a 'reasonable selection'. Anyway, already in the operationalization of the Subset there is an implicit decision regarding what variables to observe. In [65] the decision comes after a literature review, so it is only natural that the authors include all the covariates in the multiverse both as x and Z. The same approach could have not been feasible for the goal of convergence into a unique set of claims in [72] or for correct identification of the causal model [18]. Multi-model analysis [56, 92] is an interesting method to solve this controversy, since it includes all combinations to begin but then it tunes the multiverse model ("Multimodel") by removing the comparatively worse specifications. However, if the goal of the multiverse is exploratory and not conclusive, a full inclusion could be more useful than risky.
    The criterion to exclude a z variable from the possibility to work as a x in the multiverse may be disconnected to any scientific evaluation and be more practical. For example, being older could decrease bone mass but the researchers could have no practical interest in just 'revert people age', while being interested in asserting dietary advice to contrast reduction of bone mass due aging. In this case the exclusion of variable 'age' from x does not mean that age does not control the impact of observed diet in the multiverse but that the statistics (p-value, estimates) of 'age' as a regressor are not reported among the multiversal statistics, since they are not of research's interest.
    b. the second decision regards how to measure the conceptual dimension implied in the hypothesis, for example, by adoption of proxies.[2]

[2] Think about the deep metrological differences between Richter and Mercalli scales in measurement of earthquake magnitude.

3. The **Dependent** y: again, there are two approaches. In a sense, it is true that two different dependents provide 'two answers for two different questions', hence they generate two different multiverses [18]; but at the same time it also makes sense the adoption of different proxies of the same response variable if the question truly regards sensitivity of the analysis to the *adoption of a proxy*, which is a legitimate research interest that can be explored through multiversal methods.
4. The **Type** of regressions: the main issue in processing the same hypothesis under different types of regression is that even after standardisation of the variables, estimates are not always comparable if not forcing some functional form for x. In binomial regression *vs.* linear regression, to keep comparability one have to force y to assume values in the unitary interval and then estimate the coefficient on $log(x)$ and at the same time avoid the logarithmic transformation on z: there is an addition of variety of vibration in a sense but also a negation of variety in another one. However, if the research is focused on p-values and not on estimates, it is definitely worth to compute multiversal statistics for more than one type of regression.
5. The **Functional Form** (f) has analogue issues to the type of regression. It is already mentioned that sometimes it depends by the decision regarding vibration of Regression Type. It is worth to remember that the function takes as many arguments as the n of covariates, hence for any $n > 3$ the size of the whole spectrum of many alternative functional forms could be impractical to compute. For $n < 3$, the impact of an exotic change in the functional form could seriously make impossible the interpretation of the coefficients. In absence of reasons to do so, having degrees of freedom regarding f design could be an error.

Decisions on Subsetting and Regression Type impact on all of $\overleftrightarrow{y_{k_y}}$, $\overleftrightarrow{F_{k_F}}$, $\overleftrightarrow{x_{k_x}}$, and $\overleftrightarrow{Z_{k_Z}}$. Each alternative decision about an element of the specification increases the number of specifications in the multiverse.

2.2 *Janus Effect*

Given a null value β_0 for coefficient and a value of α of statistical significance, the "Janus effect" [62, 65] is the presence in the multiverse of regressor x of vibrated estimates $\hat{\beta}_x > \beta_0$ and $\hat{\beta}_x < \beta_0$, both such that $p(\beta) < \alpha$[3].

The interpretation of the presence of Janus effect is worrisome: it means that, given α, it is possible for a researcher to claim both positive or negative association between two observational variables in a population. Just by mapping a multiverse and specifying an *ad hoc* model, a desired causal claim can be p-hacked. The implications of Janus effect for clinical research are broadly discussed in [65] and [62].

[3] Janus was the Roman god of gates and was always represented with two faces pointing towards opposite directions, hence the name of the effect.

Table 1 Janus matrix of a x-multiverse

	$\beta_x < \beta_0$	$\beta_x > \beta_0$
$p(\beta_x) < \alpha$	A	B
$p(\beta_x) \geq \alpha$	C	D

Table 2 Janus confusion matrix

	One-faced	Two-faced
$p(\beta_x) < \alpha$	$\lvert A - B \rvert$	$(A + B) - \lvert A - B \rvert$
$p(\beta_x) \geq \alpha$	$\lvert C - D \rvert$	$(C + D) - \lvert A - B \rvert$

Given the dependency of Janus effect from α and β_0 is not surprising to see proposals to lower α or to shift the research on estimation of intervals for coefficients instead of looking for significant effects.

In [65] Janus effect is treated mostly as something that is there or is not, however in a multiverse made of many specifications, the magnitude of Janus effect can be observed through counting how many significant specifications hold $\beta_x > \beta_0$, and how many $\beta_x < \beta_0$.

More in general, all the estimates $\hat{\beta}_x$ in a x-multiverse can be represented through a tetrachoric matrix (Table 1):

The two dimensions in Table 1 do not share the same proprieties, though: $p(\beta_x) < \alpha$ does signal a desired condition, $p(\beta_x) \geq \alpha$ does not. The same cannot be said by comparing $\beta_x < \beta_0$ and $\beta_x > \beta_0$. The desirable outcome is to maximise into 1 the ratio:

$$\frac{|A - B|}{A + B} \tag{3}$$

which can be interpreted as the fraction of the significant results leading towards a supposedly *true* direction of the coefficient.

One can catch here the analogy of the statistical measure of Precision $\frac{TruePositives}{Positives}$. However, Precision alone does not account for sensitivity of the test to false negatives, so is usually paired to Recall $\frac{TruePositives}{True}$.

The whole Table 1 can be remapped as a Confusion Matrix (Table 2):

So, if $Precision \simeq \frac{|A-B|}{A+B}$, then,

$$Recall \simeq \frac{|A - B|}{|A - B| + |C - D|} = \frac{|A - B|}{|(A + C) - (B + D)|}. \tag{4}$$

3 Materials

3.1 Why Coronavirus

COVID-19 emerged in Wuhan (China) in late 2019. In two years, the virus has spread to more than 200 countries worldwide. The outbreak was declared a pandemic by the World Health Organisation on February 22, 2020, and hundreds of millions of COVID-19 cases have been reported, causing millions of deaths [90].

Given the uniqueness of the virus and its biological characteristics, it has been able to spread so rapidly that COVID-19 has become a public health problem [49]. The rapid growth in infection rates in each country has had a severe impact on the capability of health-care services to tackle the pandemic. As discussed in [66], many policies were implemented to reduce the deaths due to the spread of the virus, to limit the growth in the number of infected people, and overall to empower the health-care services. In addition, population control strategies have been implemented, too.

The most monitored and analysed variables associated with infection risk [69] to study the COVID-19 spread have been demographic characteristics [23], passenger mobility [12, 57], air pollution [17, 19], and comorbidity [8, 9, 26, 63].

The rapid spread of COVID-19 has triggered an uncommon increase in research activities leading to an extensive production of several observational studies. However, most of the studies do not go beyond modeling the relationship between COVID-19 and some variables (e.g., air pollution). As a consequence, a data analysis based only on single relationships could be limited and provide misleading results.

In particular, the case of the COVID-19 outbreak is showing the critical role of information dissemination which can strongly influence people's behaviour and alter the effectiveness of countermeasures implemented by governments [27, 45].

It is common for researchers to explore several analytical alternatives [11, 41, 68, 74], i.e., to look for a significant combination in order report only it [5, 48, 84]. Multiple approaches are capable of drawing causal findings from observational data as shown in [7, 53]. In all of the approaches, substantial uncertainty remains about the best model to apply. A multiverse of possible alternatives (or other forms of robustness checks) needs in order to explore how the findings would differ if different assumptions were been adopted [36].

In the present study, a multiverse modeling approach is proposed to process COVID-19 pandemic data. Multiverse analysis is proposed as a suitable method to analyse data when uncertainty could lead to mis-specification of relationships among variables [73, 74].

3.2 Selection of Covariates

One of the challenges in epidemiology is that epidemics happen within societies, and societies are very complex phenomena with a lot of features being relevant for epidemiological models.

Following the example of [62, 65], and [14], the present study aims at selecting covariates that have been broadly discussed in the epidemiological literature of COVID-19 and are available through National Public Health Departments.

The dataset of the present study is made by 16 covariates (see, Table 3). Of the 16, 3 variables are entered as dependents in the models:

1. the count of hospitalised patients with COVID-19
2. the count of hospitalised in intensive care (ICU) with COVID-19
3. the reported count of cases of COVID-19, in the countries.

The observed values of these 13 variables are summary statistics of epidemiological dimensions observed in 20 European countries. They are counted in 4 different time intervals plus the cumulative count from the start of the first interval to the end of the fourth, so, for 20 countries, the total amount of observations in the dataset is $(4 * 20) + 20 = 100$. The other 13 population variables are not collected along the 4 time intervals but they are updated at 2019 and and 2020, so their values are fixed across time.

All the variables and the counts are normalised to the population of each country and then standardised. Furthermore, four conceptual dimensions are detected as shown in Table 3:

Details about the selected variables on the basis of the more updated research findings in the literature are below described.

COVID-19 Data

Metrics on the spread of Covid over the four phases (Sect. 4) are those collected by World Health Organisation[4] and the European Center of Disease and Control (ECDC)[5], which collect case data submitted by national governments. Where possible, they aim to report confirmed cases.

The main difference among the three cumulative metrics of COVID-19 in their impact over health-care systems, is that the count of reported cases is likely to be biased by many sources of under-reporting while hospitalised patients, both in ICU and not, are unlikely to be asymptotically biased [89]. The variable $Hosp$ refers to people who have contracted COVID-19 and need hospitalization, both in the ICU or in other hospital departments and the variable ICU includes only patients in ICU departments. Both the variables deal with the two conditions (infected and in need of hospitalisation) as if two events are causally independent. These are also referred as "patients-*with*-COVID-19" and are distinct from "patients-*for*-COVID-19", which

[4] https://www.who.int/.

[5] https://www.ecdc.europa.eu/en.

Table 3 Variables in multiverse model

Dimension	Description of variable	Labels	Source	Year
COVID-19	N. Confirmed Cases	Cases	National Public Health Dept.	
COVID-19	N. Intensive Care Unit	ICU	National Public Health Dept.	
COVID-19	N. Hospitalised	Hosp	National Public Health Dept.	
Demography	Urban Population Index	UrbanPop	EuroStat	2020
Demography	N. over 65+ years	Over65	EuroStat	2020
Demography	Population Density Index	PopDensity	EuroStat	2020
Health	% Cardiovascular risk	Cardio	EuroStat	2020
Health	% Diabetes prevalence	Diabetes	EuroStat	2020
Health	% Smokers	Smoking	EuroStat	2020
Health	% Obeses	Obesity	EuroStat	2020
Health	% High blood pressure	HiPressure	EuroStat	2020
Pollution	PM2.5 Index	PM2.5	EuroStat	2019
Pollution	PM10 Index	PM10	EuroStat	2019
Pollution	CO2 Index	CO2	EuroStat	2019
Mobility	N. aeroportual passengers	AirPass	EuroStat	2019
Mobility	N. train passengers	TrainPass	EuroStat	2019

is the case when COVID-19 induces hospitalisation. ICU and $Hosp$ are two mutual proxies.

Sources of under-reporting of COVID-19 cases are due to both delays in reporting cases and prevalence of asymptomatic infected people. In particular, the sources claim that suspected cases are not be reported. Another issues dealing with different countries and institutions is that the delay in updating the number of confirmed cases is never consistent among cases. This is due to differences in times of reporting a new tested case and its inclusion in national statistics.

In general, especially in the first phase (see, Table 4), the number of confirmed cased is underestimated. Nevertheless this outcome could be a *vulnus* for scientific research about the spread of the virus, the divergence between $Cases$ and the two variables ICU and $Hosp$ is useful to shown the capability of multiversal methods in regressive analysis.

Demography

Epidemics spread over populations. All the variables in the dataset are weighted to the total population of the countries, but three variables are selected in particular to summarises demographic characteristics of the country. These are: the ratio of people aged over 65 at 2020, the index of urbanisation and the density of population.

Table 4 Epidemic phases

Phase	Start	End	What happened
1st	30/12/2020	30/03/2020	WHO reported evidence of transmission from symptomatic, pre-symptomatic, and asymptomatic infected people with COVID-19
2nd	01/04/2020	23/11/2020	The UK authorities reported a variant of SARS-CoV-2 to the WHO
3rd	24/11/2020	01/01/2021	Pfizer/Biontech vaccine was the first to receive emergency use validation from WHO for efficacy against COVID-19
4th	02/01/2021	30/03/2021	End of data collection

Most official data sources report more severe impacts of COVID-19 on the elderly [47], probably due both to an inherent weakness of their immune system and to the coexistence of other chronic diseases. According to [66], the age is a very important predictor of severe COVID-19. The risk of severe outcomes increases sharply by age, even after controlling for other potential confounding factors, including sex and pre-existing disease conditions. Age is also an important confounder in the associations between some underlying conditions and severe COVID-19 outcomes.

The index of urbanisation is the proportion of people living in a urban center over total population of the country [25]. It measures the demographic current phenomenon of population mobility from rural to urban areas. The high concentration of people and activities in urban areas makes vulnerable the populations to the exposure to COVID-19 infection due to the large amount of social networking [76].

In addition, many studies [1, 31] provide evidences on the correlation between density and the spread of the pandemic. The population density, due to economic and social reasons, affects the spread of COVID-19 infection and the incidence of the cases in the territorial areas.

Comorbidity

Comorbidity is the presence of two or more conditions occurring in a patient, either at the same time or successively. Population with multiple co-existing illness conditions are widespread [22]. This awareness has led to a growing interest among practitioners and researchers in assessing the impact of comorbidity on mortality, health-related quality of life, and efficiency of health care systems.

In the present study, population variables observed as potentially associated to patients infected with COVID-19 are selected: diabetes [26], obesity [8, 9], smoking [23], hypertension, and cardiovascular risk [64].

However, data related to above population variables are not collected from medical records of COVID-19 hospitalised patients but they are epidemiological surveillance data [40] on risk factors for the public health due to the onset or the complications of diseases.

Passengers Mobility

According to [12], mobility data are often used to correlate population mobility and the spread of an infection.

Based on the current literature, in countries where mobility is high, the number of people infected with COVID-19 is higher [11]. Likely there is a positive association between a high airport mobility in a country and a high spread of COVID-19 infection and, as a consequence, the number of both ICU and other hospital admissions increases [60].

In addition to the number of airport passengers, to capture mobility between countries and within national borders, also the number of train passengers are taken into account due both lacking of airports in some territorial areas and regular use of train for mobility.

Air Pollution

According to [19, 66], there are differences in the association between the spread of COVID-19 and the concentration levels of particulate matter (PM10 and PM2.5), and CO2. Long-term previous exposure to air pollution could be an important mediator of deaths by COVID-19 in Europe. In addition, the latest estimates made by the European Environment Agency (EEA) [24] show that the exposure to particulate matter has a strong impact on health [57, 63].

Air pollution has been postulated to affect the viability and transport of viral particles in the air. Long-term exposure could increase the risk of infection by altering the immune system [44]. Particulate matter (PM) is able to deeply enter into the respiratory tract and increase the risk of respiratory diseases [17]. Exposure to PM2.5 is positively associated with COVID-19 infection and with severity of the disease [15, 75].

Furthermore, for the purposes of the present analysis, the choice to *vibrate* the y helps to simulate the sensitivity of the linear regressive model specified on observational data to a non linear epidemic phenomenon as COVID-19 pandemic[43].

In order to define the timings of observation of COVID-19 pandemic, in Table 4 are shown 4 time phases:

Data are collected in 20 European countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Portugal, Slovenia, Spain, Sweden, United Kingdom.

All of countries provide summary statistics of the variables through to National Health Departments and Eurostat.[6]

[6] UK provides demographic data to Eurostat being a EU member until 2021. Other countries (e.g., Poland) are excluded from the dataset due to missing values.

Table 5 Correlation coefficients among variables

	Cases	ICU	Hosp
Cases	1	−0.4	−0.4
ICU	−0.4	1	0.91
Hosp	−0.4	0.91	1
UrbanPop	0.03	0.11	0.07
Over65	−0.09	0.01	0.06
PopDensity	0.12	0.10	0.05
Cardio	−0.03	−0.01	0.11
Diabetes	−0.02	−0.07	−0.12
Smoking	0.03	0.01	0.07
Obesity	−0.03	−0.11	−0.14
HiPressure	−0.04	−0.06	−0.01
PM2.5	0.15	−0.07	−0.06
PM10	0.14	−0.07	−0.06
CO2	0.10	−0.03	−0.04
AirPass	−0.09	−0.03	−0.09
TrainPass	0.01	−0.00	−0.01

4 Methods: Multiverse Models

The values of correlation coefficients (Table 5) show that in the dataset the cases of COVID-19 are unaligned with the two variables of hospital data. Only ICU and $Hosp$ are strongly correlated (0.91), since ICU is obviously a proportion of $Hosp$, hence the high correlation.

All the values of Bravais-Pearson's correlation coefficients between the 3 y and the other 13 covariates are much weaker than the correlations among the 3 y. This is the reason for *vibrating* the regressive models into multiverses in order to explore how these fixed observations are congruent with causal relationships $x \rightarrow y$ [43].

To generate the multiverse, the covariates assume the role of x or 'first regressor' in the linear model:

$$y \sim \beta_0 + \beta_x(x) + B_{K_x}(K_n \subseteq Z_x) + \epsilon \tag{5}$$

where Z_x is the set of all possible additions of all the *other non-x*, z-covariates as control, e.g., $Z_x : (K_1 = \emptyset, K_2 = z_1, K_3 = z_1 + z_2, K_4 = z_1 + z_2 + z_3, \ldots, K_n = z_2, K_n = z_2 + z_3, \ldots, K_{max(n)-1} = z_{11} + z_{12}, K_{max(n)} = z_{12})$ and B_{K_x} is the vector of coefficients associated to K_n.

$\hat{\beta}_x$ and $p(x)$ are recorded for each vibration. However, while all the non-y covariates are members of each Z_x, not all the covariates assume the role of 'first regressor' x. Indeed, the variables of the Demographic dimension ($UrbanPop$, $Over65$, and

$PopDensity$) are excluded from this role and assume only the role of z, hence their p-values and β_x are not recorded in the result.

The reason for this exclusion is twofold: the first is the fact that variables as age, density and urban density must have an impact in epidemics does not generate any kind of scientific controversy, or at least it does not ask the same scientific questions as variables as the rate of obesity or railway mobility; the second reason of exclusion is that other variables have space of intervention for a specific, clear-cut public policy (for example, an obesity reduction program), while the claim of a causal relationship between old age and an epidemic would still require other circumstantial assessments of feasibility that cannot be captured by a linear model.

No other functional form nor different scale of measurement for the variables are modeled for the multiverse with the only exception of PM10 and PM2.5, which can be considered mutual proxies since their value of Pearson's correlation $r \sim 0.99$. In this sense the adopted methodology is more akin to EBA and VoE than to Multiverse Analysis or Specification Curve. However, these differences regard more the goal of the research than the technical procedure.

4.1 Measuring the Outcome of a Shift in α

In the Sect. 2 is mentioned the proposal to shift the conventional α from 0.05 to 0.005. The multiverse offers an opportunity to evaluate the consequence of this change. The main expected result by the shift is to reduce ambiguity in causal interpretation of estimates of a regressive model. However, this result comes at the cost of a higher degree of false negatives.

The goodness of α at making much harder to reach ambiguous results (see, Sect. 2) can be measured by the pseudo-Precision in (3). To ponder this metric to the sensitivity to false negatives, pseudo-Recall (4) is measured too and the two measures are compounded into a indicator J_α as harmonic mean of the two ones (6):

$$J_\alpha = \frac{2}{\frac{A+B}{|A-B|} + \frac{|(A+C)-(B+D)|}{|A-B|}} = \frac{2\,A - B|}{A + B + |A - B + C - D|} \tag{6}$$

which is an analogue of the F_1 Score. This metric has been criticised by Chicco and Jurman [13], so the Phi correlation (ϕ_α) and tetrachoric correlation ($r_{tet,\alpha}$) on the Janus Confusion Matrix (Table 2) are provided as alternative metrics [39, 86].

Finally, to estimate the impact of the shift from $\alpha = 0.05$ to $\alpha = 0.005$, the net differences in J, ϕ, and r_{tet} are measured:

$$\begin{aligned} \Delta(J) &= J_{\alpha=0.005} - J_{\alpha=0.05} \\ \Delta(\phi) &= \phi_{\alpha=0.005} - \phi_{\alpha=0.05} \\ \Delta(r_{tet}) &= r_{tet,\alpha=0.005} - r_{tet,\alpha=0.05}. \end{aligned} \tag{7}$$

5 Results

For each y, ten linear models with a different first regressor x have been vibrated through all the possible combinations of 12 controls (plus the 'no control' case). The observations regarding y are expanded by splitting the counts among 4 time phases (see, Table 4).

This generates 409,600 specifications for each y split in 10 groups of 40,960 specifications for each first regressor, multiplying *per* 3 y, the total is of 30 groups. Multiverse statistics summarise information regarding these 30 groups. The total amount of observed specifications in the study is 409,600 * 3 = 1,228,800.

The p-curves of the 10 multiversal linear models are plotted as density curves in Figs. 1, 2, and 3.

In Fig. 1 is shown an effect between PM2.5, PM10, Air Passenger, and Cases.

The Figs. 2 and 3 illustrate that Cardiovascular risk and Smoking are very important predictors for ICU cases and Hosp cases. However, for ICU cases, Air Passengers show a significant relationship similar to Fig. 1. Instead, Obesity plays an important role in $Hosp$.

Multivariate statistics for the multiverse $x \rightarrow Cases$ are displayed in Table 6, for $x \rightarrow ICU$ in Table 7, and for $x \rightarrow Hosp$ in Table 8.

For models on COVID-19 cases (Table 6), lowering α leads into an average negative impact in the indicators of evaluation, with the exception of PM10. However, looking at the VoE plot of PM10 and PM2.5 (Fig. 4), this result is mostly determined by imbalance among classes in the Janus Confusion Matrix of PM10. Indeed, once results are paired with those of the proxy of PM2.5, significant results for $\alpha < 0.005$ are paradoxically opposite. Given also the small effect size of the estimates, these are spurious occurrences determined by overfitting of the specifications.

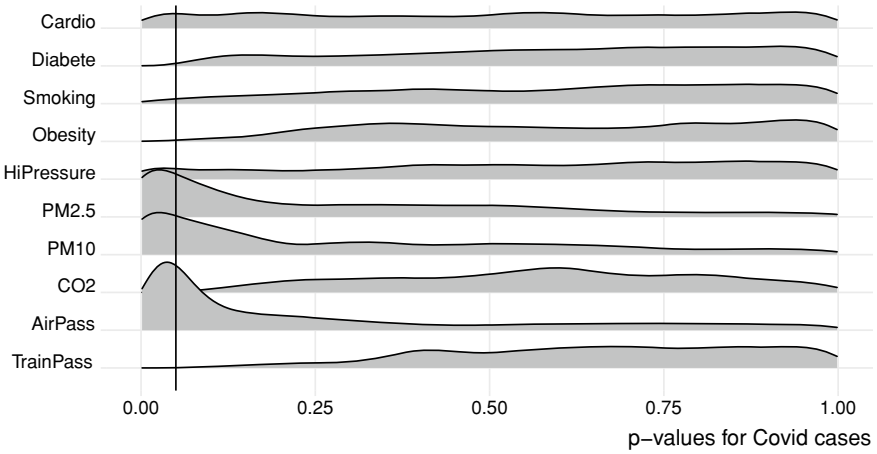


Fig. 1 Density of p-values in the multiverse $x \rightarrow Cases$

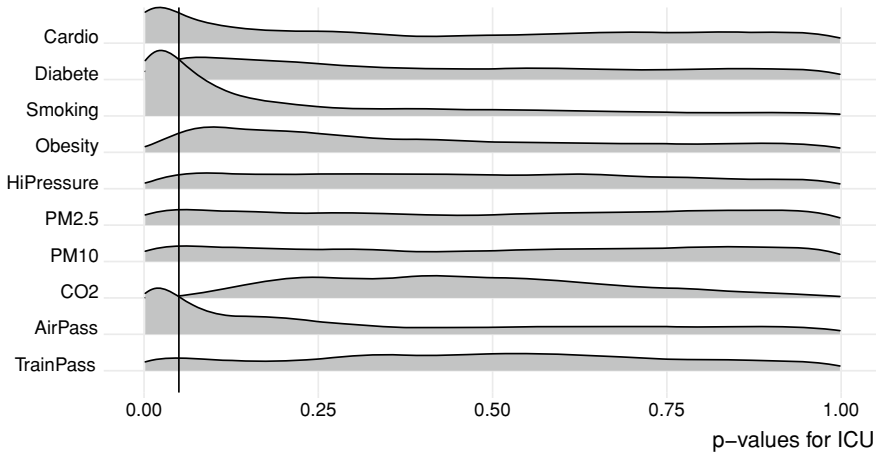


Fig. 2 Density of p-values in the multiverse $x \rightarrow ICU$

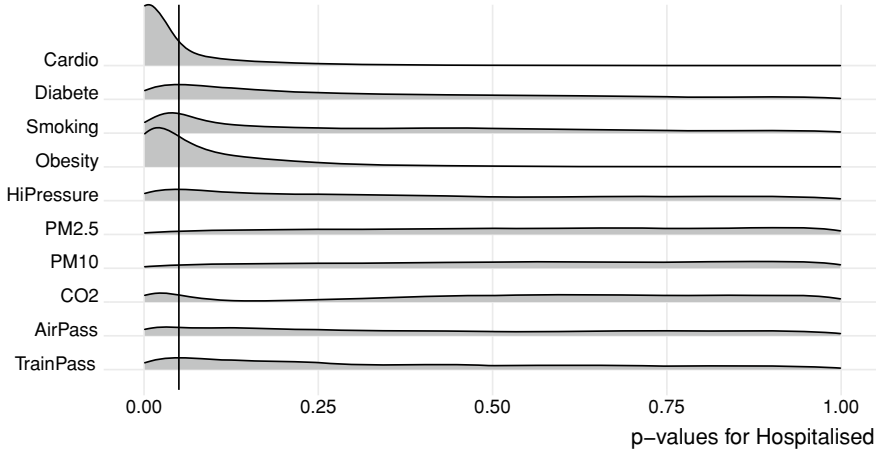


Fig. 3 Density of p-values in the multiverse $x \rightarrow Hosp$

In Fig. 4 is noteworthy the presence of a mirror Janus effect around β_0. In this case this result is reached by controlling x through its proxy, which is an easy avoidable statistical fallacy.

For models $ICU \sim x$ (Table 7), the analysis confirms that lowering the α is not helpful to highlight good scientific findings and in many cases it could be detrimental.

For example, one can consider the VoE for $ICU \sim Diabetes$, as shown in Fig. 5.

In Fig. 5 there is a clear-cut case of Janus effect that can be common in observational studies. Since $Diabetes$, unlike PM, is relatively robust to 'bad controls', the estimate span is much more narrow even if all the variables are standardised.

Table 6 Summary statistics for the multiverse $Cases \sim x$

x	r_ρ	$\%(p < 0.05)$	Med. (p)	$\bar{\beta}$	$\Delta(J)$	$\Delta(\phi)$	$\Delta(r_{tet})$
Cardio	−0.03	0.06	0.52	−0.06	−0.14	−0.11	−0.52
Diabetes	−0.02	0.00	0.63	0.06	−0.01	−0.05	−0.41
Smoking	0.03	0.01	0.63	0.06	−0.04	−0.06	−0.29
Obesity	−0.03	0.00	0.61	−0.05	−0.00	−0.03	−0.60
HiPressure	−0.04	0.05	0.59	−0.01	−0.30	−0.30	−0.17
PM2.5	0.15	0.25	0.19	−1.14	−0.26	−0.10	−0.10
PM10	0.14	0.22	0.21	−0.93	0.33	0.50	0.69
CO2	0.10	0.00	0.58	0.06	−0.00	−0.01	−0.20
AirPass	−0.09	0.31	0.12	−0.24	−0.51	−0.22	−0.54
TrainPass	0.01	0.00	0.67	0.02	0.00	0.00	0.00

Table 7 Summary statistics for the multiverse $ICU \sim x$

x	r_ρ	$\%(p < 0.05)$	Med. (p)	$\bar{\beta}$	$\Delta(J)$	$\Delta(\phi)$	$\Delta(r_{tet})$
Cardio	−0.01	0.21	0.31	0.03	−0.22	−0.02	0.30
Diabetes	−0.07	0.07	0.38	−0.07	−0.66	−0.65	−0.69
Smoking	0.01	0.38	0.09	0.05	−0.39	−0.08	−0.19
Obesity	−0.11	0.04	0.33	−0.12	−0.16	−0.22	−0.87
HiPressure	−0.06	0.04	0.45	−0.04	−0.27	−0.04	0.26
PM2.5	−0.07	0.07	0.52	−0.16	0.36	0.27	0.43
PM10	−0.07	0.07	0.51	−0.33	−0.11	0.02	0.25
CO2	−0.03	0.00	0.44	−0.03	−0.01	−0.05	−0.51
AirPass	−0.03	0.27	0.20	−0.03	−0.31	−0.16	−0.10
TrainPass	−0.00	0.06	0.49	−0.05	−0.16	−0.14	−0.22

Table 8 Summary statistics for the multiverse $Hosp \sim x$

x	r_ρ	$\%p < 0.05$	Med. (p)	$\bar{\beta}$	$\Delta(J)$	$\Delta(\phi)$	$\Delta(r_{tet})$
Cardio	0.11	0.76	0.00	0.16	−0.20	0.00	−0.17
Diabetes	−0.12	0.15	0.23	−0.15	−0.28	−0.14	−0.26
Smoking	0.07	0.21	0.24	0.08	−0.37	−0.14	−0.44
Obesity	−0.14	0.49	0.05	−0.17	−0.43	−0.06	−0.21
HiPressure	−0.01	0.12	0.33	−0.08	−0.28	−0.10	0.20
PM2.5	−0.06	0.03	0.56	0.10	−0.06	0.02	0.58
PM10	−0.06	0.03	0.56	−0.01	−0.13	−0.07	0.36
CO2	−0.04	0.10	0.58	−0.04	−0.26	−0.21	−0.33
AirPass	−0.09	0.10	0.39	−0.07	−0.21	−0.13	−0.16
TrainPass	−0.01	0.12	0.30	−0.08	−0.23	−0.11	−0.30

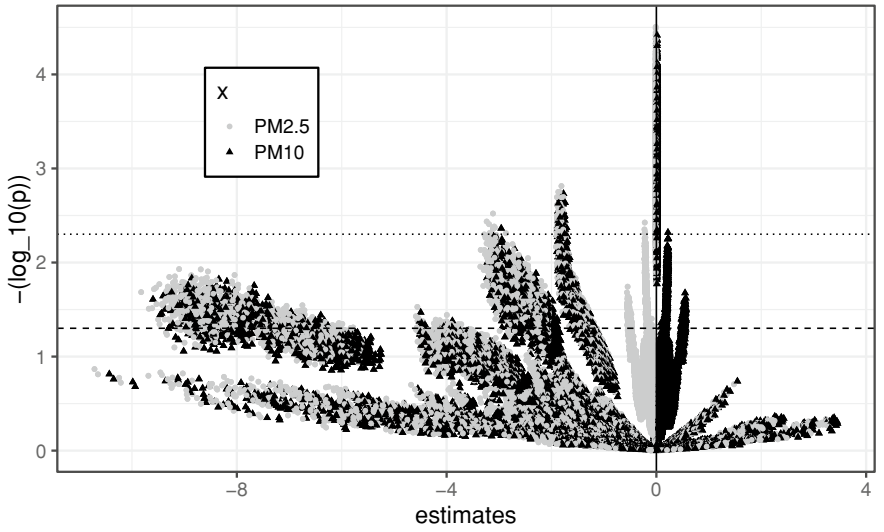


Fig. 4 VoE plot of Covid *cases* ∼ Levels of Particulate Matter

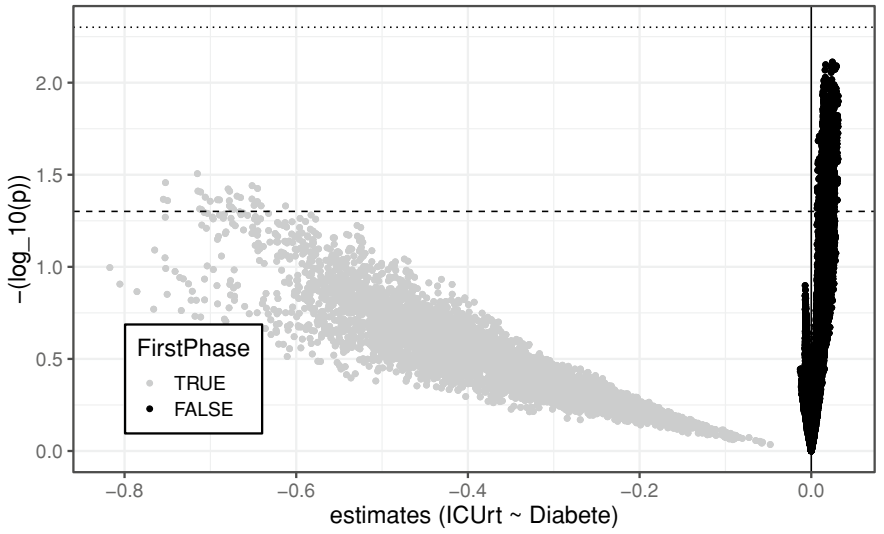


Fig. 5 VoE plot of ICU patients with Covid ∼ % of people with diabetes in the country

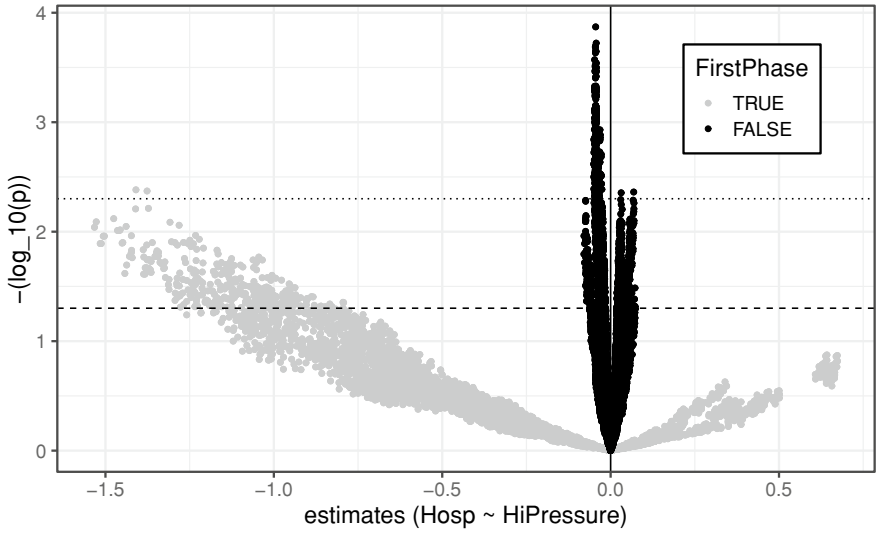


Fig. 6 VoE plot of $Hospitalised$ with COVID-19 $\sim$ % of people with high pressure in the country

However, all the estimates of specifications, when run on the data available before April 1st, 2020 (see, Table 4) could lead to think that people affected by diabetes would be, in a way or another, less prone to ICU by COVID-19 (given also a negative linear correlation between the two variables). Indeed, this is only an effect of a hasty analysis that would lead into a scientific controversy:

- $\alpha = 0.05$ is already good 'gatekeeper' since only a minority of specifications can be p-hacked into a significant result.
- Time reveals that if there is a causal relationship between diabetes and harsh-COVID-19, it is extremely weak and positive, even if $\hat{\beta}$ of the multiverse is negative, so it would be easy to p-hack.

Since $\alpha = 0.05$ is a good filter while $\alpha = 0.005$ misses the opportunity to actually 'say something about' (no specification is significant under such level), the evaluation indicators penalise $\alpha = 0.005$ so much.

For the multiverse of $Hosp$, $HiPressure$ is the most ambiguous case (Table 8, Fig. 6). For $\alpha = 0.05$ this is a clear-cut case of Janus effect. Considering the whole dataset of 100 observations, lowering the level of significance to $\alpha = 0.005$ would point towards a unique interpretation of the relationship: it is negative and it is weak. Considering only the first Phase, by $\alpha = 0.05$ while the inference of the sign would have been correct, the size of the negative effect would be overestimated.

The result would have been very hard to p-hack under a level of significance $\alpha = 0.005$. Indicators for Janus effect do not detect significant differences in estimates within the same class, they do only the relation between estimates $\hat{\beta}$ and β_0. However the tetrachoric correlation identifies in this case the potential benefit of adoption of $\alpha = 0.005$.

6 Discussion

In this study 3 indicators of Janus effect have been employed to evaluate the consequence of a shift towards the level of significance $\alpha = 0.005$ in terms of risk of mis-specification of a small-entity causal relationship in observational small samples. The impact has been estimated on a multiverse which is generated by an observational sample with high volatility in the lagged y dependent but fixed x.

$\alpha = 0.005$ does a good job of 'gatekeeping' from the possibility of p-hacking a desired outcome but this is a tautology if compared to all of the cases where there is a nearly unambiguous relation (see, Fig. 5) and the relation would be lost as a False Negative setting such a low value of α.

It could still be argued that the impact of False Negatives can be weighted by the low estimates and $\alpha = 0.005$ would not miss stronger effects. If this is the case, then is true that null hypothesis testing provides only a limited contribution to the epidemiological research on observational data and, as a consequence, methodologies focused on effect estimation and not on the statistical significance of the relationship, should be adopted.

The main limitation of the study regards the adoption of J index, ϕ index, and tetracoric correlations as indicators of the overall impact of α on the trustworthiness of claims from observational data. α is there to decrease false positives in scientific studies, so it is consequential that it is penalised by measures that evaluate the rate of false negatives.

The limitation of the Janus Confusion Matrix is that it is insensitive to the actual rate of $+$ and $-$ estimates in the multiverse but only to their absolute divergence, i.e. there could be cases where $+$ is dominant per $p(\beta_x) \geq \alpha$ and $-$ is dominant per $p(\beta_x) < \alpha$.

The above Fig. 5 highlights one of these cases. In these scenarios, the shift of α is actually effective at isolating a True scientific claim but the measures of the study could not capture it. However, occurrences like these are rare and better and more specific analytical choices in generation of the multiverse should totally avoid them [18].[7]

References

1. Almagro, M., Orane-Hutchinson, A.: JUE insight: the determinants of the differential exposure to COVID-19 in New York city and their evolution over time. J. Urban Econ. (2020). https://doi.org/10.1016/j.jue.2020.103293
2. Anderson, S.F., Maxwell, S.E.: Addressing the "replication crisis": using original studies to design replication studies with appropriate statistical power. Multivar. Behav. Res. **52**(3), 305–324 (2017)

[7] The paper is also part of research line on vulnerability and risk management of the project *GRIDAVI* Risk Management, Decision Uncertainties and Social Vulnerabilities by the University Research Incentive Plan 2020/2022 called PIACERI.

3. Athey, S., Imbens, G.: A measure of robustness to misspecification. Am. Econ. Rev. **105**(5), 476–80 (2015)
4. Begley, C.G., Ellis, L.M.: Raise standards for preclinical cancer research. Nature **483**(7391), 531–533 (2012)
5. Benjamin, D.J., et al.: Redefine statistical significance. Nat. Hum. Behav. **2**(1), 6–10 (2018)
6. Bossuyt, P.M.: Laboratory measurement's contribution to the replication and application crisis in clinical research. Clin. Chem. **65**(12), 1479–1480 (2019)
7. Bruns, S.B., Ioannidis, J.P.: P-curve and p-hacking in observational research. PLoS ONE **11**(2), e0149144 (2016)
8. Busetto, L., et al.: Obesity and COVID-19: an Italian snapshot. Obesity **28**(9), 1600–1605 (2020)
9. Caci, G., et al.: COVID-19 and obesity: dangerous liaisons. J. Clin. Med. **9**(8), 2511 (2020)
10. Camerer, C.F., et al.: Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat. Hum. Behav. **2**(9), 637–644 (2018)
11. Chang, S., et al.: Mobility network models of COVID-19 explain inequities and inform reopening. Nature **589**(7840), 82–87 (2021)
12. Charaudeau, S., et al.: Commuter mobility and the spread of infectious diseases: application to influenza in France. PLoS ONE **9**(1), e83002 (2014)
13. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. **21**(1), 1–13 (2020)
14. Chu, L., et al.: Vibration of effects in epidemiologic studies of alcohol consumption and breast cancer risk. Int. J. Epidemiol. **49**(2), 608–618 (2020)
15. Conticini, E., et al.: Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? Environ. Pollut. **261**,(2020). https://doi.org/10.1016/j.envpol.2020.114465
16. Cosme, D., et al.: Multivariate neural signatures for health neuroscience: assessing spontaneous regulation during food choice. Soc. Cogn. Affect. Neurosci. **15**(10), 1120–1134 (2020)
17. Cui, Y., et al.: Air pollution and case fatality of SARS in the People's Republic of China: an ecologic study. Environ. Health **2**(1), 1–5 (2003)
18. Del Giudice, M., Gangestad, S.W.: A traveler's guide to the multiverse: promises, pitfalls, and a framework for the evaluation of analytic decisions. Adv. Methods Pract. Psychol. Sci. **4**(1) (2021). https://doi.org/10.1177/2515245920954925
19. Donzelli, G., et al.: Relations between air quality and COVID-19 lockdown measures in Valencia, Spain. Int. J. Environ. Res. Public Health **18**(5), 2296 (2021)
20. Duvendack, M., et al.: What is meant by "replication" and why does it encounter resistance in economics? Am. Econ. Rev. **107**(5), 46–51 (2017)
21. Earp, B.D., Trafimow, D.: Replication, falsification, and the crisis of confidence in social psychology. Front. Psychol. (2015). https://doi.org/10.3389/fpsyg.2015.00621
22. Ebinger, J.E., et al.: Pre-existing traits associated with Covid-19 illness severity. PLoS ONE **15**(7), e0236240 (2020)
23. Espejo-Paeres, C., Núñez-Gil, I.J., Estrada, V., et al.: Impact of smoking on COVID-19 outcomes: a HOPE Registry subanalysis. BMJ Nutr. Prev. Health **4**, (2021). https://doi.org/10.1136/bmjnph-2021-000269
24. European Environmental Agency: Air Quality in Europe—2020 report, 9/2020, EEA Report (2020)
25. Eurostat: Eurostat Regional Yearbook, Edition 2020 (2020)
26. Fadini, G.P., et al.: Newly-diagnosed diabetes and admission hyperglycemia predict COVID-19 severity by aggravating respiratory deterioration. Diabetes Res. Clin. Pract. **168** (2020). https://doi.org/10.1016/j.diabres.2020.108374
27. Gauchat, G.: Politicization of science in the public sphere: a study of public trust in the United States, 1974 to 2010. Am. Sociol. Rev. **77**(2), 167–187 (2012)
28. Gelman, A., Loken, E.: The statistical crisis in science. Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. Am. Sci. **102**(6), 460–466 (2014)

29. Granger, C.W., Uhlig, H.F.: Reasonable extreme-bounds analysis. J. Econometrics **44**(1–2), 159–170 (1990)
30. Halsey, L.G., et al.: The fickle P value generates irreproducible results. Nat. Methods **12**(3), 179–185 (2015)
31. Hamidi, S., et al.: Does density aggravate the COVID-19 pandemic? Early findings and lessons for planners. J. Am. Plann. Assoc. **86**(4), 495–509 (2020)
32. Harder, J.A.: The multiverse of methods: extending the multiverse analysis to address data-collection decisions. Perspect. Psychol. Sci. **15**(5), 1158–1177 (2020)
33. Head, M.L., et al.: The extent and consequences of p-hacking in science. PLoS Biol. **13**(3) (2015). https://doi.org/10.1371/journal.pbio.1002106
34. Hicks, D.J.: Open science, the replication crisis, and environmental public health. Accountability Res. 1-29 (2021). https://doi.org/10.1080/08989621.2021.1962713
35. Imbens, G.W.: Statistical significance, p-values, and the reporting of uncertainty. J. Econ. Perspect. **35**(3), 157–74 (2021)
36. Ioannidis, J.P., et al.: Increasing value and reducing waste in research design, conduct, and analysis. The Lancet **383**(9912), 166–175 (2014)
37. Ioannidis, J.P.: The proposal to lower P value thresholds to 0.005. JAMA **319**(14), 1429–1430 (2018)
38. Ioannidis, J.P.: What have we (not) learnt from millions of scientific papers with P values? Am. Stat. **73**(sup1), 20–25 (2019)
39. Islam, T.U., Rizwan, M.: Comparison of correlation measures for nominal data. Commun. Stat. Simul. Comput. 1–20 (2020). https://doi.org/10.1080/03610918.2020.1869984
40. ISS EpiCentro: I dati per l'Italia. La Sorveglianza Passi d'Argento (2020). https://www.epicentro.iss.it/passi-argento
41. Jewell, N.P., et al.: Predictive mathematical models of the COVID-19 pandemic: underlying principles and value of projections. JAMA **323**(19), 1893–1894 (2020)
42. Johnson, V.E.: Revised standards for statistical evidence. Proc. Natl. Acad. Sci. **110**(48), 19313–19317 (2013)
43. Klau, S., et al.: Sampling uncertainty versus method uncertainty: a general framework with applications to omics biomarker selection. Biometrical J. **62**(3), 670–687 (2020)
44. Kogevinas, M., et al.: Ambient air pollution in relation to SARS-CoV-2 infection, antibody response, and COVID-19 disease: a cohort study in Catalonia, Spain (COVICAT study). Environ. Health Perspect. **129**(11) (2021). https://doi.org/10.1289/EHP9726
45. Kreps, S.E., Kriner, D.L.: Model uncertainty, political contestation, and public trust in science: evidence from the COVID-19 pandemic. Sci. Adv. **6**(43) (2020). https://doi.org/10.1126/sciadv.abd4563
46. Leamer, E.E.: Sensitivity analyses would help. Am. Econ. Rev. **75**(3), 308–313 (1985)
47. Lee, Y.J.: The impact of the COVID-19 pandemic on vulnerable older adults in the United States. J. Gerontol. Soc. Work **63**(6–7), 559–564 (2020)
48. Leek, J.T., Peng, R.D.: Statistics: P values are just the tip of the iceberg. Nat. News **520**(7549), 612 (2015)
49. Mao, Z., et al.: Investigating the self-reported health status of domestic and overseas Chinese populations during the COVID-19 pandemic. Int. J. Environ. Res. Public Health **18**(6), 3043 (2021)
50. Masur, P.K.: Understanding the effects of conceptual and analytical choices on 'finding' the privacy paradox: a specification curve analysis of large-scale survey data. Inf. Commun. Soc. 1–19 (2021). https://doi.org/10.1080/1369118X.2021.1963460
51. Masur, P.K., Scharkow, M.: specr: Conducting and Visualizing Specification Curve Analyses. R Package (2020)
52. Maxwell, S.E., et al.: Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am. Psychol. **70**(6), 487–498 (2015)
53. Mayo, D.G., Spanos, A.: Methodology in practice: statistical misspecification testing. Philos. Sci. **71**(5), 1007–1025 (2004)

54. Mazzola, J.J., Deuling, J.K.: Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I-O journal articles. Ind. Organ. Psychol. **6**(3), 279–284 (2013)
55. McShane, B.B., et al.: Abandon statistical significance. Am. Stat. **73**(sup1), 235–245 (2019)
56. Muñoz, J., Young, C.: We ran 9 billion regressions: eliminating false positives through computational model robustness. Sociol. Methodol. **48**(1), 1–33 (2018)
57. Nižetić, S.: Impact of coronavirus (COVID-19) pandemic on air transport mobility, energy, and environment: a case study. Int. J. Energy Res. **44**(13), 10953–10961 (2020)
58. Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251) (2015). https://doi.org/10.1126/science.aac4716
59. Orben, A., Przybylski, A.K.: The association between adolescent well-being and digital technology use. Nat. Hum. Behav. **3**(2), 173–182 (2019)
60. Oztig, L.I., Askin, O.E.: Human mobility and coronavirus disease 2019 (COVID-19): a negative binomial regression analysis. Public Health **185**, 364–367 (2020). https://doi.org/10.1016/j.puhe.2020.07.002
61. Page, L., et al.: The replication crisis, the rise of new research practices and what it means for experimental economics. J. Econ. Sci. Assoc. 1–16 (2021). https://doi.org/10.1007/s40881-021-00107-7
62. Palpacuer, C., et al.: Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. BMC Med. **17**(1), 1–13 (2019)
63. Pansini, R., Fornacca, D.: COVID-19 higher mortality in Chinese Regions with chronic exposure to lower air quality. Front. Public Health **8**, (2021). https://doi.org/10.3389/fpubh.2020.597753
64. Parohan, M., et al.: Risk factors for mortality in patients with Coronavirus disease 2019 (COVID-19) infection: a systematic review and meta-analysis of observational studies. The Aging Male **23**(5), 1416–1424 (2020)
65. Patel, C.J., et al.: Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J. Clin. Epidemiol. **68**(9), 1046–1058 (2015)
66. Perone, G.: The determinants of COVID-19 case fatality rate (CFR) in the Italian regions and provinces: an analysis of environmental, demographic, and healthcare factors. Sci. Total Environ. **760**, (2021). https://doi.org/10.1016/j.scitotenv.2020.142523
67. Pike, H.: Statistical significance should be abandoned, say scientists. BMJ **364**, (2019). https://doi.org/10.1136/bmj.l1374
68. Pluchino, A., et al.: A novel methodology for epidemic risk assessment of COVID-19 outbreak. Sci. Rep. **11**(1), 1–20 (2021)
69. Qiu, F., et al.: Impacts of cigarette smoking on immune responsiveness: up and down or upside down? Oncotarget **8**(1), 268–284 (2017)
70. Rohrer, J.M., et al.: Probing birth-order effects on narrow traits using specification-curve analysis. Psychol. Sci. **28**(12), 1821–1832 (2017)
71. Sala-i-Martin, X.: I just ran four million regressions. Am. Econ. Rev. **87**(2), 178–183 (1997)
72. Saraceno, J., et al.: Reevaluating the substantive representation of lesbian, gay, and bisexual Americans: a multiverse analysis. J. Politics **83**(4), 1837–1843 (2021)
73. Schmeiser, H., et al.: The risk of model misspecification and its impact on solvency measurement in the insurance sector. J. Risk Finance **13**(4), 285–308 (2012)
74. Seitshiro, M.B., Mashele, H.P.: Quantification of model risk that is caused by model misspecification. J. Appl. Stat. 1–21 (2020). https://doi.org/10.1080/02664763.2020.1849055
75. Setti, L., et al.: Potential role of particulate matter in the spreading of COVID-19 in Northern Italy: first observational study based on initial epidemic diffusion. BMJ Open **10**(9), e039338 (2020)
76. Sharifi, A., Khavarian-Garmsir, A.R.: The COVID-19 pandemic: impacts on cities and major lessons for urban planning, design, and management. Sci. Total Environ. (2020). https://doi.org/10.1016/j.scitotenv.2020.142391
77. Simmons, J.P., et al.: False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. **22**(11), 1359–1366 (2011)

78. Simonsohn, U., et al.: Specification curve analysis. Nat. Hum. Behav. **4**(11), 1208–1214 (2020)
79. Simonsohn, U., et al.: Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available at:SSRN (2019). https://doi.org/10.2139/ssrn.2694998
80. Sönning, L., Werner, V.: The replication crisis, scientific revolutions, and linguistics. Linguistics **59**(5), 1179–1206 (2021)
81. Steegen, S., et al.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. **11**(5), 702–712 (2016)
82. Taroni, F., et al.: Statistical hypothesis testing and common misinterpretations: should we abandon p-value in forensic science applications? Forensic Sci. Int. **259**, e32–e36 (2016)
83. Trafimow, D.: Five nonobvious changes in editorial practice for editors and reviewers to consider when evaluating submissions in a post $p < 0.05$ universe. Am. Stat. **73**(sup1), 340–345 (2019)
84. Trafimow, D., et al.: Manipulating the alpha level cannot cure significance testing. Front. Psychol. **9**, (2018). https://doi.org/10.3389/fpsyg.2018.00699
85. Vanpaemel, W., et al.: Are we wasting a good crisis? The availability of psychological research data after the storm. Collabra **1**(1), 1–5 (2015)
86. Warrens, M.J.: On association coefficients for $2\times$ 2 tables and properties that do not depend on the marginal distributions. Psychometrika **73**(4), 777 (2008)
87. Wasserstein, R.L., Lazar, N.A.: The ASA statement on p-values: context, process, and purpose. Am. Stat. **70**(2), 129–133 (2016)
88. Wasserstein, R.L. et al.: Moving to a world beyond "$p < 0.05$". Am. Stat. **73**(sup. 1), 1–19 (2019)
89. Wei, E.K., et al.: Nine lessons learned from the COVID-19 pandemic for improving hospital care and health care delivery. JAMA Intern. Med. **181**(9), 1161–1163 (2021)
90. World Health Organization: Coronavirus Disease 2019 (COVID-19) (2021). Available at: https://apps.who.int/iris/handle/10665/331475
91. Yaffe, J.: From the editor—do we have a replication crisis in social work research? J. Soc. Work Educ. **55**(1), 1–4 (2019)
92. Young, C., Holsteen, K.: Model uncertainty and robustness: a computational framework for multimodel analysis. Sociol. Methods Res. **46**(1), 3–40 (2017)

Neural Network for the Statistical Process Control of HVAC Systems in Passenger Rail Vehicles

Fiorenzo Ambrosino, Giuseppe Giannini, Antonio Lepore, Biagio Palumbo, and Gianluca Sposito

Abstract In the rail industry, coach temperature regulation has become a crucial task to improve passenger thermal comfort. Over the past few years, European standards have required rail operators to implement monitoring systems for the control of heating, ventilation and air conditioning (HVAC) of passenger rail vehicles. These systems, based on modern automated sensing technologies, have created new data-rich scenarios and call for new methods to deal with high-dimensional, high-correlated and heterogeneous data. In this article, an autoencoder, which is a particular type of neural network developed to model unlabelled data and automatically extract significant features, is utilised to develop a nonparametric process monitoring approach. Two control charts based on statistics H^2 and SPE are built in the feature space and the residual space, respectively. Through operational HVAC data collected on board passenger vehicles, the proposed approach is shown to be capable of simultaneously monitoring and detecting anomalies that may have occurred in the data streams acquired from each train coach, even though it is not limited to the application hereby investigated. Additionally, via a numerical investigation, the Phase II fault detection performance is compared with that of a simpler linear dimension reduction method and two more complex NN architectures.

Keywords Statistical process control · Autoencoder · Railway HVAC systems · Multivariate control chart

F. Ambrosino
ENEA, Italian National Agency for New Technologies Energy and Sustainable Economic Development, Portici, Italy

G. Giannini
Head of Operation Service and Maintenance Product Evolution, Hitachi Rail STS, Naples, Italy

A. Lepore · B. Palumbo · G. Sposito (✉)
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy
e-mail: gianluca.sposito@unina.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_23

1 Introduction

In the rail industry, efficient temperature regulation is becoming a necessity in the face of strong competition and overcrowded carriages. One of the challenging aspects in this regard is the passenger thermal comfort of rail coaches, especially for long journeys. Over the past few years, new European standards have been developed, such as UNI EN 14750 [1], and prescribe requirements for controlling the air temperature, relative humidity and air speed within passenger compartments. In order to meet these standards, railway companies have been automatically collecting and storing data for the monitoring of heating, ventilation and air conditioning (HVAC) systems installed on board modern trains.

During the working life of a system, different kinds of faults may occur and compromise the originally designed functions. The timely identification of faults in a complex system is in fact a crucial task to ensure operation safety and quality as well as to reduce maintenance and operation costs. The challenge is to turn the collected high dimensional, correlated and heterogeneous data, also possibly affected by noise and environmental fluctuations, into value. The ongoing digitalization creates in fact a new dimension in the diagnosis, maintenance and operation of these systems, in a new cost-effective and efficient manner. In this modern data-rich and computationally efficient environment, classical statistical process control (SPC) [2] and machine learning (ML) methods [3] interplay without sharp boundaries. Traditional methods, no matter if based on model-driven [4] or data-driven [5] approaches, may sometimes achieve poor performance, due to the increasing complexity of the data produced by modern industrial processes. As a result, alternative monitoring methods based on neural networks (NNs), which in the past decades have been regarded as unaffordable because computationally expensive and large data demanding, have recently received a great attention [6–8] also in the field of SPC [9, 10].

In particular, the class of supervised ML methods [3] are naturally prone to automatically identify normal and out-of-normal system behaviours, but require to be trained based on large labelled data sets with, ideally, the same number of data for each behaviour. Unfortunately, this is not feasible in practice, because systems generally work under normal conditions for the most part of their life and faults are generally very rare, making real data sets unbalanced in this sense. On the other hand, unsupervised ML approaches do not require data with labels or additional information, but have the drawback of showing lower performances for fault detection and classification. Hence, in this perspective, we explore the use of autoencoders (AEs) [11], which are a particular type of NN capable of extracting from unlabelled data significant features that are not necessarily linearly related to the original inputs. AE output will be exploited to be used in a nonparametric SPC charting scheme and mitigate the drawbacks of both supervised and unsupervised ML methods. In this way, we possibly avoid artificially changing normal system operating conditions to produce out-of-normal behaviours, and we can model the raw data signals directly, without the need for arbitrarily selecting and extracting problem-specific features based on past experience, if any.

The rest of the paper is organised as follows. In Sect. 2, the proposed approach is introduced. In Sect. 3, it is applied to the operational data acquired for the monitoring of HVAC systems installed on board passenger railway vehicles, and made available by the rail transport company Hitachi Rail STS. Finally, in Sect. 4, theoretical and practical conclusions are drawn. An Appendix closes the paper by comparing the Phase II fault detection performance of the proposed method with that of a simpler linear dimension reduction method, and two more complex NN architectures, on four different simulated non-normal operating condition scenarios.

2 The Proposed Approach

In the following subsections, the proposed SPC approach is introduced by briefly framing NNs (Sect. 2.1), and, in particular, AEs (Sect. 2.2). We then present (Sect. 2.3) how to use the typical AE output for the perspective monitoring of new data points, usually referred to as Phase II in the SPC literature. It is then worth readily stating that to start Phase II, we must preliminary identify a clean set of observations that represents the process normal operating conditions and will be hereinafter referred to as *reference set* or *training set*. The identification of the training set is called Phase I, and cyclically applies offline the following steps: (i) the training of the AE and hyperparameter optimisation on an initial, as large as possible, set of observations; (ii) the application of the SPC approach based on the AE output; (iii) the recognition of points that fall out the control limits; (iv) the possible elimination of outlying points marked as exceptional by domain experts; (v) and the revision of the control limits [2]. For conciseness, and without loss of generality, we assume in the following subsections that a training set has been already identified and available. As a standard practice in NN literature, step (v) is performed without repeating step (i), and, in particular, hyperparameter optimisation in step (i) is performed by means of a validation set or k-fold cross-validation approach [3] on the training set, which consists of N vectors denoted by $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, N$.

2.1 Artificial Neural Networks

A NN is a brain-inspired system, as the adjective *neural* suggests, with the goal of replicating the human learning process. NNs are usually introduced as advanced algorithms capable of extracting complex and nonlinear relationships among variables and are a popular tool used for learning and visualisation problems, such as computer vision, speech recognition, natural language processing and function approximation. A NN consists of a collection of interconnected processing elements, called *neurons*, which loosely mimic the biological neuron in a human brain. Each neuron receives input from other neurons, then processes and transmits it to the neurons connected to it. Typically neurons are organised in *layers*, which are divided into three types:

input, *hidden* and *output*. The input layer is used to bring the data to the network without performing any computations and passing the information to the hidden layer. The hidden layer, placed between input and output layers, is responsible for the relationship mapping among input variables in the model. More than one hidden layer can exist, depending on the complexity of the training set. The output layer, in turn, takes inputs from the hidden layer and returns the final output. Based on the type of connections among neurons, two successive layers are called *fully connected*, if each neuron of a layer is connected to each neuron in the successive layer, or *pooling*, if a group of neurons in one layer is connected to a single neuron in the successive layer. NNs with these types of connections are called *feedforward NN*, and the information moves only in one direction from the input layer to the output layer without forming any cycle. When feedforward NNs are extended to include connections between neurons in the same or previous layers, they are called *recurrent* NNs.

2.2 Autoencoder

An AE is a type of NN typically employed for dimensionality reduction and feature extraction with the goal of copying the input to the output, possibly getting the significant aspects of the data to be copied. It usually consists of two symmetric parts: an encoder and a decoder [11]. The encoder maps each input vector $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, into $\mathbf{h}_i = f(\mathbf{x}_i) \in \mathbb{R}^{d'}$, called code, or latent, representation stored in its last layer, called code layer. Starting from the latter, the decoder maps the code representation back to a reconstructed vector or output, in the input space, denoted by $\mathbf{x}'_i = r(\mathbf{h}_i) = r(f(\mathbf{x}_i)) \in \mathbb{R}^d$. AEs have one input layer, one output layer and an appropriate number of hidden layers, each with an optimal number of neurons. In particular, the output layer must have the same number of neurons as the input layer, as they respectively belong to the decoder and encoder, which are symmetric, being the code layer the interface between the encoder and the decoder. However, perfectly copying the input to the output may be useless, whereas the goal is to summarise, into the code layer only the input properties that are important for the problem at hand and discard the others. To achieve this, one may force $\mathbf{h}_i$ to have a dimension that is smaller than that of the input. An AE whose architecture is subject to this constraint is called *undercomplete*. Learning an undercomplete representation forces the AE to separate important information from noise. If the hidden layer dimension is larger than or equal to that of the input, the AE is called *overcomplete* and may merely mimic the identity function. There exist various strategies for preventing this [11]. The proposed approach is based on an undercomplete, feedforward and fully connected AE with only one hidden layer, for which $\mathbf{h}_i$ is computed for each input $\mathbf{x}_i$ as follows

$$\mathbf{h}_i = g(\mathbf{W}\mathbf{x}_i + \mathbf{b}). \tag{1}$$

The function $g(\cdot)$ is called the *activation function*. $\mathbf{W}$ and $\mathbf{b}$ are the $d' \times d$ *weight matrix* and the *bias vector*, respectively, and are referred to the neuron connections

between the input layer and the hidden layer. The most common activation functions are the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$, the hyperbolic tangent (tanh) $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, the rectified linear unit (ReLU) $g(z) = \max(0, z)$ and the exponential linear unit (ELU), which returns the input itself if it is positive, and $\gamma(e^z - 1)$ otherwise, $\gamma \in [0, 1]$. Then, $\mathbf{x}_i'$ can be accordingly computed as for each $i = 1, \ldots, N$

$$\mathbf{x}_i' = g'(\mathbf{W}'\mathbf{h}_i + \mathbf{b}'), \tag{2}$$

where $\mathbf{W}'$ and $g'(\cdot)$ are the $d \times d'$ weight matrix and the activation function of the decoder, respectively, and $\mathbf{b}'$ is the bias vector referred to the neuron connections between output layer and hidden layer. AE training (step (i)) is performed by minimising the mean squared error of all reconstruction errors $\mathbf{e}_i$ between input $\mathbf{x}_i$ and its reconstruction $\mathbf{x}_i'$. The objective function to be minimised is defined as follows

$$\mathrm{J}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x}_i, r(f(\mathbf{x}_i))) = \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x}_i, \mathbf{x}_i'), \tag{3}$$

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{W}', \mathbf{b}, \mathbf{b}'\}$ denotes the training parameter vector, $L(\cdot, \cdot)$ is referred to as the *loss function* and penalises $\mathbf{x}_i'$ for being dissimilar from the input vector $\mathbf{x}_i$, i.e., it is increasing with the reconstruction error $\mathbf{e}_i = \mathbf{x}_i - \mathbf{x}_i'$. In this paper, for each i, $L(\mathbf{x}_i, \mathbf{x}_i')$ is assumed to be the sum of squared errors and therefore, it can be explicitly written as the squared L_2-norm of $\mathbf{e}_i$

$$\begin{aligned} L(\mathbf{x}_i, \mathbf{x}_i') &= \|\mathbf{x}_i - \mathbf{x}_i'\|^2 = \|\mathbf{e}_i\|^2 \\ &= \|\mathbf{x}_i - g'(\mathbf{W}'(g(\mathbf{W}\mathbf{x}_i + \mathbf{b})) + \mathbf{b}'\|^2. \end{aligned} \tag{4}$$

Coherently, we denote by $\boldsymbol{\theta}_{\boldsymbol{o}} = \{\mathbf{W}_\mathbf{o}, \mathbf{W}_\mathbf{o}', \mathbf{b}_\mathbf{o}, \mathbf{b}_\mathbf{o}'\}$ the optimal parameter vector that results from the minimisation of the objective function obtained from Eqs. (3) and (2.2), as follows

$$\boldsymbol{\theta}_{\boldsymbol{o}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \|\mathbf{x}_i - g'(\mathbf{W}'(g(\mathbf{W}\mathbf{x}_i + \mathbf{b}))) + \mathbf{b}'\|^2. \tag{5}$$

This problem can be solved by means of the backpropagation algorithm [12], where the gradient descent approach [11] is used to update the parameter vector $\boldsymbol{\theta}^t$ at each iteration t as follows

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \epsilon(\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{t-1})) \tag{6}$$

where ϵ is called learning rate and $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{t-1})$ is the gradient of the cost function $L(\boldsymbol{\theta}^{t-1})$ with respect to the weight matrices and the bias vectors at iteration $t-1$ for each layer. The weight matrices and the bias vectors are initialised by means of the Xavier initialisation algorithm [13] and updated until an early stopping criterion [11, 14] is met to finally get $\boldsymbol{\theta}_{\boldsymbol{o}}$. Specifically, the training will be stopped when the loss

function of the NN does not improve on a hold out validation set for 10 consecutive epochs. The pseudocode of the AE training algorithm is given in Algorithm 1.

Algorithm 1 AE training algorithm

Input: Reference set
Output: encoder and decoder NNs
Initialise the parameter vector $\boldsymbol{\theta}$ by using Xavier initialisation
repeat
  Sample a minibatch from the reference set
  Compute the loss function $L(\mathbf{x}_i, r(f(\mathbf{x}_i)))$
  Update $\boldsymbol{\theta}$ by gradient descent approach
until the early stopping criterion does not meet

When the encoder and decoder have linear activation functions and the mean squared error is used as a loss function, an undercomplete AE performs the well-known principal component analysis (PCA) [15]. Conversely, when a nonlinear activation function is used, AEs may learn to span a different subspace from that identified by PCA [16]. In this perspective, AEs generalise the PCA and are expected to achieve better performance in latent feature extraction, projection, and classification [17].

It can be proved for example that a NN with one hidden layer can approximate any continuous function from the input patterns to the output patterns with an arbitrary degree of accuracy [11], provided that there are enough neurons in the hidden layer. Anyway, the selection of a proper number of neurons in the hidden layer, the type of activation function, the learning rate and the batch size is a delicate task belonging to the hyperparameter optimisation of step (i).

2.3 *Construction and Control Limit Estimation of the Monitoring Statistics*

The output of the AE from step (i) can be embedded in a control charting scheme and enhance a yet familiar SPC tool for practitioners. Implementation of process monitoring based on AEs is similar to those based on PCA [18]. Specifically, the monitoring statistics H^2 [19] and the squared prediction error (SPE) [20] are developed to improve the identification of possible faults that may have occurred in the process. For each input $\mathbf{x}_i$, the H^2 statistic [19] is defined in the feature space as the squared L_2-norm of the latent representation $\mathbf{h}_i$ obtained from Eq. (1), that is

$$H_i^2 = \mathbf{h}_i^\top \mathbf{h}_i = (g(\mathbf{W}\mathbf{x}_i + \mathbf{b}))^\top g(\mathbf{W}\mathbf{x}_i + \mathbf{b}), \quad i = 1, \ldots, N. \tag{7}$$

The H^2 statistic is the square distance of the projection of the input $\mathbf{x}_i$ from the origin of the feature space [20]. Whereas, the SPE statistic [20] is defined in the residual space as the squared L_2-norm of the reconstruction error, explicitly written upon

using Eq. (2) as follows

$$\begin{aligned} SPE_i &= \mathbf{e}_i^\top \mathbf{e}_i = (\mathbf{x}_i - \mathbf{x}_i')^\top (\mathbf{x}_i - \mathbf{x}_i') \\ &= (\mathbf{x}_i - g'(\mathbf{W}'\mathbf{h}_i + \mathbf{b}')^\top (\mathbf{x}_i - g'(\mathbf{W}'\mathbf{h}_i + \mathbf{b}')), \quad i = 1, \ldots, N. \end{aligned} \tag{8}$$

The SPE statistic measures how close the input $\mathbf{x}_i$ is to the residual space by the squared perpendicular distance of $\mathbf{x}_i$ from the feature space [20]. These two monitoring statistics can be then plotted against a time index and form two different control charts named H^2 and SPE control charts and can be used by practitioners to visualise any process drift (i.e., trends) and to issue an alarm if an observation of at least one monitoring statistic falls above the relative upper control limit (UCL). In particular, the H^2 control chart monitors the variation inside the feature space spanned by the feature extracted by the AE, whereas changes along directions orthogonal to the latter space are monitored by the SPE control chart, which signals faults that make the input to move away from the feature space defined by the reference model. Therefore, it is not surprising that the H^2 statistic is affected by a type of fault that does not increase the value of the SPE statistic, or viceversa. The UCLs of the H^2 and SPE control charts, denoted by H^2_{lim} and SPE_{lim}, respectively, are estimated using the block bootstrap approach [21–24]. In particular, these will be computed by taking an average of the $(1-\alpha')$-quantiles of the H^2 and SPE statistics computed on each of the 1000 bootstrap samples that are drawn from the reference set. The block bootstrap approach is an alternative to the traditional kernel density estimation (KDE) procedure when the observations of a given statistic are autocorrelated, and thus the assumption of independent and identically distributed data, which KDE relies on [25], does not hold. The block bootstrap method is instead proven to allow estimation of a sampling distribution of a statistic even for strongly dependent sequences [21–23]. Note that, to ensure a *family wise error rate* (FWER) [26] smaller than or equal to the desired threshold α, α' is chosen by using the Sidák correction [26] $\alpha' = 1 - (1-\alpha)^{1/2}$. The control charts are then ready to be used in Phase II. Let $\mathbf{x}_{\text{new}}$ denote a new data point and $\mathbf{h}_{\text{new}} = g(\mathbf{W}\mathbf{x}_{\text{new}} + \mathbf{b})$ to denote its representation obtained through the trained encoder network. Then, the corresponding value of H^2 and SPE to be reported in the relative control chart, can be calculated by Eqs. (7) and (8), respectively, upon substituting $\mathbf{x}_i$ with $\mathbf{x}_{\text{new}}$ and $\mathbf{h}_i$ with $\mathbf{h}_{\text{new}}$.

3 Real-Case Study

HVAC data collected on board modern passenger vehicles, courtesy of Hitachi Rail STS, is employed to demonstrate the effectiveness of the proposed approach. Section 3.1 introduces the HVAC data and the available variables used to define the model. In Sect. 3.2, the main results are presented and discussed.

Table 1 Operational variables measured for each train's coach

Variable	Description
T_{Return}	Interior temperature
T_{Outdoor}	Exterior temperature
T_{SetPoint}	Target temperature set by European regulations
T_{Supply}	Theoretical temperature provided by the HVAC

3.1 HVAC and Data Description

Real operational data are collected every two minutes for about two months in the summer season from modern HVAC systems installed onboard a passenger 6-coach rail vehicle. Table 1 summarises the 4 variables related to the HVAC operating conditions that are available for each of the 6 coaches and used for both training and testing purposes. That is, the dimension of each input observations $\mathbf{x_i}$ is equal to $d = 24$. The training set contains $N = 45776$ samples drawn from the process under normal conditions. Whereas the test set used for Phase II contains 504 samples, and it is known that the process mean shifts starting from sample 176. Train names and voyage dates are intentionally omitted for confidentiality reasons.

A central unit is installed to control the HVAC system performance through temperature sensors that activate heating or cooling mode based on measurements of outside (T_{Outdoor}) and interior (T_{Return}) temperatures. The aim of the central unit is to maintain a comfortable temperature and humidity levels throughout the range of environmental and climatic conditions required by the aforementioned European regulations UNI EN 14750-1 [1]. In particular, the HVAC central unit implements [1] by dynamically setting, as a function of T_{Return} and T_{Outdoor}, the target temperature (T_{SetPoint}) that is the desired temperature value at which the HVAC systems attempts to maintain the T_{Return} value at each time instant. Actually, [1] allows $|T_{\text{Return}} - T_{\text{SetPoint}}|$ to be no larger than 2°C, otherwise the train cannot be operational.

3.2 Results

The number of input and output layer nodes are both set equal to the dimension of the input observation, that is $d = 24$. The remaining optimal AE hyperparameters are chosen by a numerical grid search as those achieving the smallest reconstruction error estimated by means of a 10-fold cross-validation applied to the training data with early stopping [11], according to what is stated in Sect. 2. Within the grid search, for each combination of the hyperparameters in the range reported in Table 2, different AEs are trained with different activation functions in each layer. Each network is trained on a single core of an Intel Xeon Platinum 8160 node of the ENEA CRESCO6 system (2.10 GHz, 192 GB RAM, no GPU) [27]. The pro-

Table 2 AE hyperparameter values. In bold, the hyperparameters, with their ranges, chosen through grid-search tenfold cross-validation procedure

Hyperparameter	Explored values	Chosen value
Learning rate	{0.1, 0.01, 0.001, 0.0001}	0.001
Number of input layer nodes		24
Number of code layer nodes	{1, 4, 10, 12, 14, 16, 18}	12
Number of output layer nodes		24
Activation function for hidden layer	{ReLU, ELU, tanh, sigmoid}	ELU
Activation function for output layer	{ReLU, ELU, tanh, sigmoid}	ELU
Batch size	{64, 128, 256, 512}	128

posed approach is numerically implemented by means of the open source software environment Python [28] and Keras [29] with TensorFlow [30] as backend. We then propose the use of a very simple AE with a single hidden layer having two neurons and parameters $\mathbf{W_o}$, $\mathbf{W'_o}$, $\mathbf{b_o}$, $\mathbf{b'_o}$ trained by backpropagation and Adam's optimisation [11] algorithms. UCLs of the monitoring statistics are estimated by means of the block bootstrap with $\alpha = 0.05$. The use of the block bootstrap is motivated in Sect. 2.3 to mitigate estimation bias due to any autocorrelation in H^2 and SPE observations in the training set.

To demonstrate the applicability of the proposed approach, Figs. 1 and 2 report the H^2 and SPE control charts based on the single hidden layer AE for the HVAC data, respectively. The dashed vertical line indicates the instant at which a fault is known to occur, whereas the bold line indicates the UCLs, H^2_{lim} and SPE_{lim}, for each chart with $\alpha' = 0.025$. After the failure of the HVAC system, SPE_i statistics fall above SPE_{lim}. Whereas, H^2_i observations, even though showing an unusual trend, do not exceed H^2_{lim}. Hence, the fault is successfully identified as a change in the residual space, and not as a drift in the feature space.

Moreover, in the Appendix, the Phase II fault detection performance of the proposed method is compared with that of a simpler linear dimension reduction method, and two more complex NN architectures, in four different scenarios, which simulate non-normal operating conditions. Different performance are obtained by setting the latent representation dimension of all methods equal to 2 and 12. In the former case the proposed method is the best in all scenarios, whereas, still being underperformed by PCA, it is the best only in detecting small deviation from normal behaviours, but can be outperformed by more complex NN architectures. It is however clear that in all cases simpler linear dimensionality reduction techniques underperforms NN-based SPC approaches.

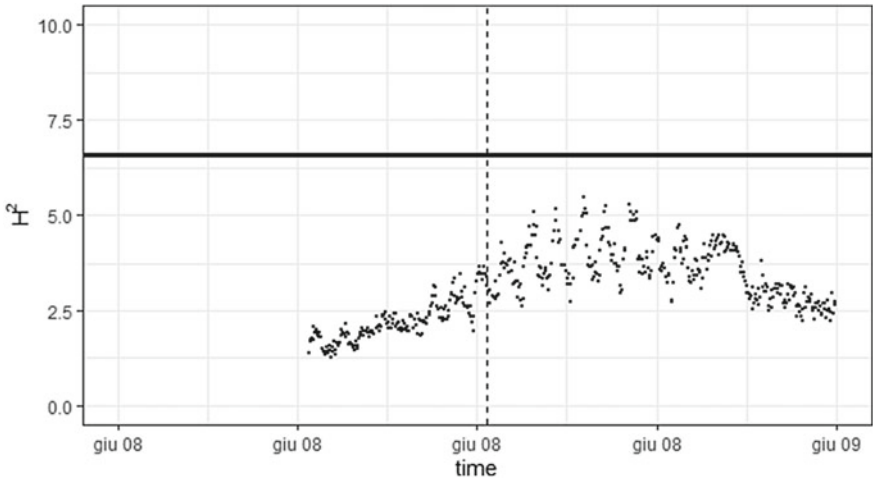


Fig. 1 H^2 control chart used for the perspective monitoring of HVAC data. Each point indicates the monitoring statistic value at each point in time. The bold line indicates the UCL at $\alpha' = 0.025$, whereas the dashed line indicates the instant at which a fault is known to occur at

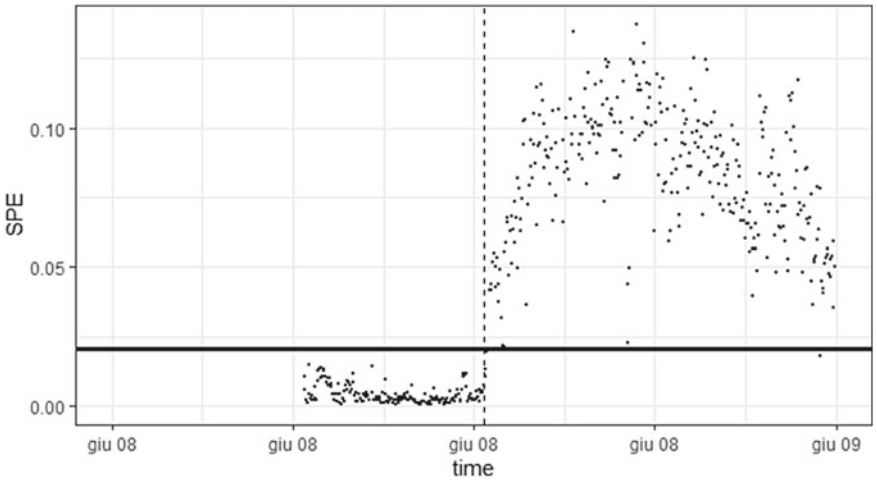


Fig. 2 SPE control chart used for the perspective monitoring of HVAC data. Each point indicates the monitoring statistic value at each point in time. The bold line indicates the UCL at $\alpha' = 0.025$, whereas the dashed line indicates the instant which a fault is known to occur at

4 Conclusions

A nonparametric statistical process control approach is proposed by means of a properly designed autoencoder (AE) neural network (NN), and the joint use of the H^2 and SPE control charts, which are built on the space spanned by the feature extracted by the AE and the corresponding reconstruction error, respectively. Phase II fault detection performance of the proposed method has been investigated by means of four simulated scenarios by setting two latent representation dimensions and has been compared with the Variational AE, a deeper AE and principal component analysis in terms of the fault detection rate (FDR) (see Appendix). To allow for a fair comparison all the comparing methods have the same number of code layer nodes. The results showed that the proposed monitoring strategy performs better than the competitors (i) in all simulated non-normal scenarios when the latent representation dimension is set to 2, and (ii) in detecting small deviations from normal behaviours with respect to more complex NN architectures when the latent representation dimension is set to 12. The approach is also shown to be capable of exploiting the massive operational HVAC data collected by the rail transport company Hitachi Rail STS and promptly indicating if and when an anomaly occurs in the HVAC system performance, in a new automated and interpretable way.

In this direction, for future work, different and possibly larger real-data examples are expected to provide further evidence of this approach benefits and applicability and to possibly be able to adapt time series approaches as Long Short-Term Memory AE.

Acknowledgements This work has been developed in the framework of the R&D project of the multiregional investment programme "REINForce: REsearch to INspire the Future" (CDS000609) with Hitachi Rail Italy, supported by the Italian Ministry for Economic Development (MISE) through the Invitalia agency. The authors are extremely grateful to engineers Vincenzo Criscuolo and Guido Cesaro, from the Operation Service and Maintenance Product Evolution Department of Hitachi Rail STS, for their technological insights in the interpretation of results. The computing resources and the related technical support used for this work have been provided by CRESCO/ENEAGRID High Performance Computing infrastructure and its staff [27]. CRESCO/ENEAGRID High Performance Computing infrastructure is funded by ENEA, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development and by Italian and European research programmes, see http://www.cresco.enea.it/english for information.

Appendix

In this Appendix, the Phase II fault detection performance of the proposed method is compared with that of PCA, AE with more than one hidden layer, referred to as DeepAE, and Variational AE, referred to as VAE [6, 11]. The PCA is regarded as a simpler linear dimension reduction method and is used as the benchmark to prove the effectiveness of non-linear approaches. Whereas DeepAE and VAE are regarded as alternative to the proposed NN with more a complex architecture.

Note that, to allow for a fair comparison, the proposed AE, the DeepAE and the VAE must be chosen with a number of code layer nodes equal to the number of components retained in the PCA. In this regard, we compare the performance of the proposed AE with that of the competing methods by setting the code layer dimension hyperparameter equal to 2 and 12. The former turned out indeed to be the optimal number of extracted features by PCA, as they explained the 82% of the total variance [16]. Whereas, the latter is the optimal choice for the number of code layer nodes for the proposed AE, as already reported in Table 2. Also in this case, as stated in Sect. 2.3 for the proposed AE, the H^2 and SPE statistics and the relative UCLs for all competing methods are estimated by means of the block bootstrap with $\alpha = 0.05$. The Phase II fault detection performance is explored by means of a numerical investigation based on four different Phase II scenarios. The latter is based on data generated by resampling and transforming $N_s = 10000$ observations from the training set in order to simulate Phase II non-normal operating conditions with the following severity levels:

1. the means of the variables T_{Return} and T_{Supply} of one coach, say coach # 1, are shifted by an amount Δ equal to two units of the relative standard deviation, i.e., $\Delta = 2\sigma$ (Scenario 1);
2. the means of the variables T_{Return} and T_{Supply} of one coach are shifted by $\Delta = 3\sigma$ (Scenario 2);
3. the means of the variables T_{Return} and T_{Supply} of two coaches, say coach # 1 and # 2, are shifted by $\Delta = 2\sigma$ (Scenario 3);
4. the means of the variables T_{Return} and T_{Supply} of two coaches are shifted by $\Delta = 3\sigma$ (Scenario 4).

Trivially, if either the H^2 or SPE statistics exceeds the value of the corresponding UCL, the fault is counted as successfully detected, otherwise, no alarm is signalled. Then, the Phase II performance is measured through the FDR index, calculated as FDR = $\text{TN}/N_s \times 100\%$, where TN is the number of fault samples correctly identified. FDRs of the four competing methods are reported in Tables 3 and 4 for a number of code layer nodes equal to 2 and 12, respectively.

In particular, Table 3 demonstrates that, even when we choose the best latent representation for the PCA, the latter is outperformed by the proposed method, which in fact performs better also than DeepAE and VAE, in all considered scenarios.

Table 3 FDR (%) for a code layer dimension hyperparameter equal to 2 in the four scenarios described in Sect. 2.3. The best performance is highlighted in bold

Scenarios	AE	PCA	DeepAE	VAE
1	**18.29**	17.72	17.81	15.32
2	**78.40**	77.10	77.16	64.58
3	**60.28**	59.84	60.12	47.44
4	**99.98**	99.97	**99.98**	99.53

Table 4 FDR (%) for a code layer dimension hyperparameter equal to 12 in the four scenarios described in Sect. 2.3. The best performance is highlighted in bold

Scenarios	AE	PCA	DeepAE	VAE
1	**16.40**	8.88	12.19	15.73
2	50.60	23.89	25.25	**58.7**
3	82.88	26	**94.67**	40.41
4	99.83	86.22	**99.98**	99.11

From Table 4, when the code layer dimension is set equal to 12, the Phase II performance of the proposed AE, although outperforming the PCA in all scenarios, is instead outperformed by VAE in Scenario 2 and DeepAE in Scenario 3 and 4. This indicates that the proposed AE is still the best in detecting small deviations from normal behaviours with respect to more complex NN architectures. It is however clear that in all cases PCA underperforms all competing non-linear approaches.

References

1. UNI-EN 14750-1: Railway Applications—Air Conditioning for Urban and Suburban Rolling Stock. Part 1: Comfort Parameters. British Standard. British Standards Institution, London (2006)
2. Montgomery, D.C.: Statistical Quality Control. Wiley Global Education (2012)
3. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 112. Springer (2013)
4. Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N.: A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Comput. Chem. Eng. **27**(3), 293–311 (2003)
5. Bersimis, S., Psarakis, S., Panaretos, J.: Multivariate statistical process control charts: an overview. Qual. Reliab. Eng. Int. **23**(5), 517–543 (2007)
6. Zhang, Z., Jiang, T., Zhan, C., Yang, Y.: Gaussian feature learning based on variational autoencoder for improving nonlinear process monitoring. J. Process Control **75**, 136–155 (2019)
7. Lee, S., Kwak, M., Tsui, K.-L., Kim, S.B.: Process monitoring using variational autoencoder for high-dimensional nonlinear processes. Eng. Appl. Artif. Intell. **83**, 13–27 (2019)
8. Chen, L., Wang, Z.-Y., Qin, W.-L., Ma, J.: Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. Signal Process. **130**, 377–388 (2017)
9. Zorriassatine, F., Tannock, J.D.T.: A review of neural networks for statistical process control. J. Intell. Manuf. **9**(3), 209–224 (1998)
10. Psarakis, S.: The use of neural networks in statistical process control charts. Qual. Reliab. Eng. Int. **27**(5), 641–650 (2011)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256. JMLR Workshop and Conference Proceedings (2010)

14. Prechelt, L.: Early stopping-but when? In: Neural Networks: Tricks of the Trade, pp. 55–69. Springer (1998)
15. Baldi, P., Hornik, K.: Neural networks and principal component analysis: learning from examples without local minima. Neural Netw. **2**(1), 53–58 (1989)
16. Jolliffe, I.: Principal Component Analysis. Encyclopedia of Statistics in Behavioral Science (2005)
17. Japkowicz, N., Hanson, S.J., Gluck, M.A.: Nonlinear autoassociation is not equivalent to PCA. Neural Comput. **12**(3), 531–545 (2000)
18. Jackson, J.E., Mudholkar, G.S.: Control procedures for residuals associated with principal component analysis. Technometrics **21**(3), 341–349 (1979)
19. Yan, W., Guo, P., Li, Z., et al.: Nonlinear and robust statistical process monitoring based on variant autoencoders. Chemom. Intell. Lab. Syst. **158**, 31–40 (2016)
20. MacGregor, J.F., Kourti, T.: Statistical process control of multivariate processes. Control Eng. Pract. **3**(3), 403–414 (1995)
21. Lahiri, S.N.: Theoretical comparisons of block bootstrap methods. Ann. Stat. 386–404 (1999)
22. Bühlmann, P., Künsch, H.R.: Block length selection in the bootstrap for time series. Comput. Stat. Data Anal. **31**(3), 295–310 (1999)
23. Härdle, W., Horowitz, J., Kreiss, J.-P.: Bootstrap methods for time series. Int. Stat. Rev. **71**(2), 435–459 (2003)
24. Phaladiganon, P., Kim, S.B., Chen, V.C.P., Baek, J.-G., Park, S.-K.: Bootstrap-based t2 multivariate control charts. Commun. Stat. Simul. Comput. ® **40**(5), 645–662 (2011)
25. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Routledge (2018)
26. Lehmann, E.L., Romano, J.P., Casella, G.: Testing Statistical Hypotheses, vol. 3. Springer (2005)
27. Iannone, F., Ambrosino, F., Bracco, G., De Rosa, M., Funel, A., Guarnieri, G., Migliori, S., Palombi, F., Ponti, G., Santomauro, G., Procacci, P.: CRESCO ENEA HPC clusters: a working example of a multifabric GPFS spectrum scale layout. In: 2019 International Conference on High Performance Computing Simulation (HPCS), pp. 1051–1052 (2019)
28. Van Rossum, G., Drake Jr., F.L.: Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam (1995)
29. Chollet, F.: Keras. https://keras.io (2015)
30. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from www.tensorflow.org (2015)

On the Use of the Matrix-Variate Tail-Inflated Normal Distribution for Parsimonious Mixture Modeling

Salvatore D. Tomarchio, Antonio Punzo, and Luca Bagnato

Abstract Recent advances in the matrix-variate model-based clustering literature have shown the growing interest for this kind of data modelization. In this framework, finite mixture models constitute a powerful clustering technique, despite the fact that they tend to suffer from overparameterization problems because of the high number of parameters to be estimated. To cope with this issue, parsimonious matrix-variate normal mixtures have been recently proposed in the literature. However, for many real phenomena, the tails of the mixture components of such models are lighter than required, with a direct effect on the corresponding fitting results. Thus, in this paper we introduce a family of 196 parsimonious mixture models based on the matrix-variate tail-inflated normal distribution, an elliptical heavy-tailed generalization of the matrix-variate normal distribution. Parsimony is reached by applying the well-known eigen-decomposition of the component scale matrices, as well as by allowing the tailedness parameters of the mixture components to be tied across groups. An AECM algorithm for parameter estimation is presented. The proposed models are then fitted to simulated and real data. Comparisons with parsimonious matrix-variate normal mixtures are also provided.

S. D. Tomarchio (✉) · A. Punzo
Università degli Studi di Catania, Dipartimento di Economia e Impresa, Catania, Italia
e-mail: daniele.tomarchio@unict.it

L. Bagnato
Università Cattolica del Sacro Cuore, Dipartimento di Scienze Economiche e Sociali, Piacenza, Italia

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_24

1 Introduction

Matrix-variate (or three-way) data have been increasingly analyzed and discussed in the statistical literature over the recent years. This data structure arises when a set of p variables is observed in r different occasions on N statistical units, thus producing a $p \times r$ matrix for each unit. Within the model-based clustering framework, several finite mixture models have been proposed for modelling this kind of data [5, 7, 13, 14, 19–23].

One serious concern related to matrix-variate mixture models is the potentially high number of parameters involved. Specifically, this issue is mainly related to the dimensionality of the two scale matrices of each mixture component, given that $p(p+1)/2 + r(r+1)/2 - 1$ unique parameters must be estimated. One effective way of addressing this problem would be to impose various constraints on the component scale matrices, thus modelling and explaining the data with as few parameters as possible. For this reason, [17] have recently generalized to the matrix-variate framework the popular class of parsimonious models proposed by [3], which are obtained by using the eigen-decomposition of the component covariance matrices. Specifically, [17] introduced a family of parsimonious mixtures based on the matrix-variate normal (MVN) distribution. However, the tails of the MVN distribution in each mixture component might not be heavy enough to adequately model many real data configurations, and mixture models based on matrix-variate heavy-tailed distributions often result in better fitting and clustering performances [5, 19, 20]. Nevertheless, these heavy-tailed matrix-variate mixture models have been proposed in an unconstrained setting, thus inheriting the above mentioned overparameterization issues. For this reason, in this paper we provide an example of parsimonious modelling using a matrix-variate heavy-tailed distribution. Specifically, we use for the mixture components the matrix-variate tail-inflated normal (MVTIN) distribution recently introduced in [19]. As in [17], we use the eigen-decomposition to attain parsimony in the component scale matrices. Furthermore, we consider the option of constraining the tailedness parameters to be tied across the mixture components. This option, together with the constraints on the scale matrices, produces the family of 196 parsimonious MVTIN mixture models (MVTIN-Ms) presented in Sect. 2, along with an alternating expectation-conditional maximization (AECM) algorithm [15] for parameter estimation.

Section 3 considers different numerical experiments. In detail, we first analyze the parameter recovery of our AECM algorithm and the capability of the Bayesian information criterion (BIC) [18] to detect the parsimonious structure of the data generating model. Then, we apply our models to a real dataset, along with parsimonious mixtures of matrix-normal distributions (MVN-Ms). The fitting performances of the competing models are compared, and the estimated parameters as well as the detected partition of the overall best fitting model are commented. Finally, some conclusions are drawn in Sect. 4.

2 Methodology

2.1 The Matrix-Variate Tail-Inflated Normal Distribution

The MVTIN distribution offers a convenient way of modelling matrix-variate data [19]. Let $\mathcal{X}$ be a continuous random matrix of dimension $p \times r$. The probability density function (pdf) of $\mathcal{X}$ following a $p \times r$-dimensional MVTIN distribution is given by

$$f_{\text{MVTIN}}(\mathbf{X}; \boldsymbol{\Theta}) = \left[\frac{2}{\delta(\mathbf{X}; \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})}\right]^{\frac{pr}{2}+1} \frac{(2\pi)^{-\frac{pr}{2}} |\boldsymbol{\Sigma}|^{-\frac{r}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}}}{\theta} \left[\Gamma\left(\frac{pr}{2}+1, (1-\theta)\frac{\delta(\mathbf{X}; \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})}{2}\right) - \Gamma\left(\frac{pr}{2}+1, \frac{\delta(\mathbf{X}; \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})}{2}\right)\right], \quad (1)$$

where $\delta(\mathbf{X}; \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right]$, $\mathbf{M}$ is the $p \times r$ mean matrix, $\boldsymbol{\Sigma}$ is the $p \times p$ row scale matrix, $\boldsymbol{\Psi}$ is the $r \times r$ column scale matrix, $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function, and $\theta \in (0, 1)$ is the tailedness parameter. $\boldsymbol{\Theta}$ contains all the parameters of the density. Regarding the tailedness parameter, we have that the closer is θ to 1, the heavier are the tails of the MVTIN distribution.

An interesting characteristic of the pdf in (1) is that it can be obtained by using the matrix-variate normal scale mixture model [8], when the mixing random variable W is uniformly distributed over the interval $(1-\theta, 1)$. This leads to the following hierarchical representation of the MVTIN distribution, that is useful for the implementation of the AECM algorithm discussed in Sect. 2.3:

1. $W \sim \mathcal{U}(1-\theta, 1)$,
2. $\mathbf{X}|W = w \sim \mathcal{N}_{p\times r}(\mathbf{M}, \boldsymbol{\Sigma}/w, \boldsymbol{\Psi})$,

where $\mathcal{U}(1-\theta, 1)$ is a uniform distribution on $(1-\theta, 1)$ and $\mathcal{N}_{p\times r}(\mathbf{M}, \boldsymbol{\Sigma}/w, \boldsymbol{\Psi})$ denotes the matrix-variate normal distribution.

Lastly, it is worth mentioning that there is a non-identifiability issue related to the two scale matrices, which are identifiable up to a multiplicative constant. Indeed, $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^* \otimes \boldsymbol{\Sigma}^*$ if $\boldsymbol{\Sigma}^* = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^* = a^{-1}\boldsymbol{\Psi}$. This is a well-known issue related to matrix-variate normal scale mixture models, for which several solutions have been proposed in the literature [7, 13, 17, 20, 23]. All the suggested solutions produce the same $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ [20], but herein we adopt the approach proposed by [13], since it is computationally more convenient. Specifically, [13] addressed this non-identifiability issue by imposing $|\boldsymbol{\Psi}| = 1$.

2.2 Parsimonious Mixtures of Matrix-Variate Tail-Inflated Normal Distributions

A $\mathcal{X}$ continuous random matrix arises from a $p \times r$-dimensional finite mixture model if its pdf can be written as

$$f_{\text{MIXT}}(\mathbf{X}; \boldsymbol{\Omega}) = \sum_{g=1}^{G} \pi_g f(\mathbf{X}; \boldsymbol{\Theta}_g), \tag{2}$$

where $f(\mathbf{X}; \boldsymbol{\Theta}_g)$ is the pdf of the gth mixture component with parameter $\boldsymbol{\Theta}_g$, π_g is the gth mixture weight, such that $0 < \pi_g \leq 1$, $g = 1, \ldots, G$ and $\sum_{g=1}^{G} \pi_g = 1$, and $\boldsymbol{\Omega}$ is the set containing all the parameters of the mixture. When the pdf in (1) is used in (2), we obtain the MVTIN mixtures proposed by [19].

However, as discussed in Sect. 1, the mixture model in (2) may be concerned by a potentially high number of parameters. To address this issue, we use the well-known eigen-decomposition of the components scale matrices. In detail, a q-dimensional scale matrix can be decomposed as

$$\boldsymbol{\Phi}_g = \lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g', \tag{3}$$

where $\lambda_g = |\boldsymbol{\Phi}_g|^{1/q}$, $\boldsymbol{\Gamma}_g$ is a $q \times q$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{\Phi}_g$, and $\boldsymbol{\Delta}_g$ is the scaled ($|\boldsymbol{\Delta}_g| = 1$) diagonal matrix of the eigenvalues of $\boldsymbol{\Phi}_g$. From a geometric point of view, λ_g indicates the volume, $\boldsymbol{\Gamma}_g$ determines the orientation, and $\boldsymbol{\Delta}_g$ represents the shape of the gth cluster. By imposing constraints on the three components of (3), we obtain the 14 parsimonious structures reported in Table 1. Since in (3) there are two scale matrices for each mixture component, this would produce $14 \times 14 = 196$ parsimonious MVTIN mixtures. However, the condition $|\boldsymbol{\Psi}| = 1$ implies that in (3) the parameter λ_g associated with the volume is unnecessary. This reduces the number of parsimonious structures related to $\boldsymbol{\Psi}$ from 14 to 7: II, EI, VI, EE, VE, EV, VV. Therefore, we have $14 \times 7 = 98$ different parsimonious structures related to the component scale matrices.

In addition to parsimony on the scale matrices, we also attain parsimony in the component tailedness parameters. Specifically, we consider the option of constraining $\theta_1, \ldots, \theta_G$ to be tied across groups. The nomenclature used for the tailedness parameter is closely related to that adopted for the scale matrices, i.e. "E" refers to tied tailedness parameters across groups and "V" is used in the unconstrained case. This option, combined with the constraints on the scale matrices, yields to a total of $98 \times 2 = 196$ parsimonious MVTIN mixtures. The nomenclature used for each parsimonious model is obtained by combining those of the scale matrices and tailedness parameter.

Table 1 Nomenclature, scale matrix structure, and number of free parameters in $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_G$ for the parsimonious models obtained via the eigen-decomposition of the component scale matrices. $\boldsymbol{I}$ is the identity matrix

Family	Model	Type	Volume	Shape	Orientation	# of free parameters in $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_G$
Spherical	EII	$\lambda \boldsymbol{I}$	Equal	Spherical	–	1
Spherical	VII	$\lambda_g \boldsymbol{I}$	Variable	Spherical	–	G
Diagonal	EEI	$\lambda \boldsymbol{\Delta}$	Equal	Equal	Axis-aligned	q
Diagonal	VEI	$\lambda_g \boldsymbol{\Delta}$	Variable	Equal	Axis-aligned	$G+q-1$
Diagonal	EVI	$\lambda \boldsymbol{\Delta}_g$	Equal	Variable	Axis-aligned	$G(q-1)+1$
Diagonal	VVI	$\lambda_g \boldsymbol{\Delta}_g$	Variable	Variable	Axis-aligned	Gq
General	EEE	$\lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	Equal	Equal	Equal	$q(q+1)/2$
General	VEE	$\lambda_g \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	Variable	Equal	Equal	$q(q+1)/2+G-1$
General	EVE	$\lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}_g\boldsymbol{\Gamma}'$	Equal	Variable	Equal	$q(q-1)/2+G(q-1)+1$
General	VVE	$\lambda_g \boldsymbol{\Gamma}\boldsymbol{\Delta}_g\boldsymbol{\Gamma}'$	Variable	Variable	Equal	$q(q-1)/2+Gq$
General	EEV	$\lambda \boldsymbol{\Gamma}_g\boldsymbol{\Delta}\boldsymbol{\Gamma}_g'$	Equal	Equal	Variable	$Gq(q-1)/2+q$
General	VEV	$\lambda_g \boldsymbol{\Gamma}_g\boldsymbol{\Delta}\boldsymbol{\Gamma}_g'$	Variable	Equal	Variable	$Gq(q-1)/2+G+q-1$
General	EVV	$\lambda \boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	Equal	Variable	Variable	$Gq(q+1)/2-G+1$
General	VVV	$\lambda_g \boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	Variable	Variable	Variable	$Gq(q+1)/2$

2.3 Parameter Estimation

We estimate the parameters of model (2) by using maximum likelihood. Within this paradigm, we implemented an AECM algorithm, which is a variant of the well-known expectation-maximization algorithm [4]. Let $\mathbf{S} = \{\mathbf{X}_i\}_{i=1}^N$ be a sample of statistical units. In the framework of our AECM algorithm, $\mathbf{S}$ is considered incomplete since we have two sources of incompleteness. In detail, the complete data are $\mathbf{S}_c = \{\mathbf{X}_i, \mathbf{z}_i, w_i\}_{i=1}^N$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})'$, such that $z_{ig} = 1$ if observation i belongs to group g and $z_{ig} = 0$ otherwise, governs the first source of incompleteness, and w_i is the realization of the mixing random variable W defined in Sect. 2.1, and is related to the second source of incompleteness. Therefore, the complete-data log-likelihood function for model (2) can be factorized as

$$\ell_c\left(\boldsymbol{\Omega}; \mathbf{S}_c\right) = \ell_{1c}\left(\boldsymbol{\pi}; \mathbf{S}_c\right) + \ell_{2c}\left(\boldsymbol{\Xi}; \mathbf{S}_c\right) + \ell_{3c}\left(\boldsymbol{\theta}; \mathbf{S}_c\right), \tag{4}$$

where

$$\ell_{1c}\left(\boldsymbol{\pi}; \mathbf{S}_c\right) = \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig} \ln\left(\pi_g\right),$$

with $\boldsymbol{\pi} = \left\{\pi_g\right\}_{g=1}^G$,

$$\ell_{2c}\left(\boldsymbol{\Xi};\mathbf{S}_c\right)=\sum_{i=1}^{N}\sum_{g=1}^{G}z_{ig}\Bigg[-\frac{pr}{2}\ln\left(2\pi\right)+\frac{pr}{2}\ln\left(w_{ig}\right)-\frac{r}{2}\ln\left|\boldsymbol{\Sigma}_g\right|-\frac{p}{2}\ln\left|\boldsymbol{\Psi}_g\right| -\frac{w_{ig}\delta(\mathbf{X};\mathbf{M}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Psi}_g)}{2}\Bigg],$$

with $\boldsymbol{\Xi}=\left\{\mathbf{M}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Psi}_g\right\}_{g=1}^{G}$ and

$$\ell_{3c}\left(\boldsymbol{\theta};\mathbf{S}_c\right)=\sum_{i=1}^{N}\sum_{g=1}^{G}z_{ig}\left\{-\ln\left(\theta_g\right)+\ln\left[\mathbb{1}_{(1-\theta_g,1)}\left(w_{ig}\right)\right]\right\},$$

where $\boldsymbol{\theta}=\left\{\theta_g\right\}_{g=1}^{G}$, and $\mathbb{1}_A\left(\cdot\right)$ is the indicator function on the set A.

Our AECM algorithm, that iteratively alternates an E-step and four CM-steps until convergence, proceeds as follows. The parameters marked with one dot correspond to the updates of the previous iteration whereas those marked with two dots represent the updates at the current iteration. Note that, standard errors for the estimates of the parameters may be operationally computed as described in [10, 11].

2.3.1 E-Step

At the E-step we have to compute

$$\ddot{z}_{ig}=\frac{\dot{\pi}_g f_{\text{MVTIN}}\left(\mathbf{X}_i;\dot{\boldsymbol{\Theta}}_g\right)}{\sum_{h=1}^{G}\dot{\pi}_h f_{\text{MVTIN}}\left(\mathbf{X}_i;\dot{\boldsymbol{\Theta}}_h\right)}\quad\text{and}$$

$$\ddot{w}_{ig}=\frac{2}{\delta(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g)}\times\frac{\Gamma\left[\frac{pr}{2}+2,\left(1-\dot{\theta}_g\right)\frac{\delta(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g)}{2}\right]-\Gamma\left[\frac{pr}{2}+2,\frac{\delta(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g)}{2}\right]}{\Gamma\left[\frac{pr}{2}+1,\left(1-\dot{\theta}_g\right)\frac{\delta(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g)}{2}\right]-\Gamma\left[\frac{pr}{2}+1,\frac{\delta(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g)}{2}\right]},$$

which correspond to the posterior probability that $\mathbf{X}_i$ belongs to the gth component of the mixture, and to the expected value of a doubly-truncated gamma distribution having parameters $(pr/2)+1$ and $\delta(\mathbf{X}_i;\mathbf{M}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Psi}_g)/2$, on the interval $\left(1-\theta_g,1\right)$, respectively [19]. Now, consider $\boldsymbol{\Omega}_1=\left\{\pi_g,\mathbf{M}_g,\boldsymbol{\Sigma}_g\right\}_{g=1}^{G}$, $\boldsymbol{\Omega}_2=\left\{\boldsymbol{\Psi}_g\right\}_{g=1}^{G}$ and $\boldsymbol{\Omega}_3=\left\{\theta_g\right\}_{g=1}^{G}$.

2.3.2 CM-Step 1

At the first CM-step we update the following parameters

$$\ddot{\pi}_g = \frac{\sum_{i=1}^{N} \ddot{z}_{ig}}{N} \quad \text{and} \quad \ddot{\mathbf{M}}_g = \frac{\sum_{i=1}^{N} \ddot{z}_{ig}\ddot{w}_{ig}\mathbf{X}_i}{\sum_{i=1}^{N} \ddot{z}_{ig}\ddot{w}_{ig}}.$$

2.3.3 CM-Step 2

At the second CM-step, keeping fixed $\boldsymbol{\Omega}_2$ at $\dot{\boldsymbol{\Omega}}_2$, we update $\boldsymbol{\Sigma}_g$. Such update depends on the parsimonious structure considered. With the exclusion of the EVE, VVE, EEV and VEV models, we limit to report only the updates of each parsimonious model, without going too deep into the discussion for the sake of brevity. Further details can be found in [3, 17]. The updates for the 14 parsimonious structures of $\boldsymbol{\Sigma}_g$ are reported in Appendix A.

2.3.4 CM-Step 3

At the third CM-step, keeping fixed $\boldsymbol{\Omega}_1$ at $\ddot{\boldsymbol{\Omega}}_1$, we update $\boldsymbol{\Psi}_g$. Also in this case, such update depends on the parsimonious structure considered. With the exclusion of the VE and EV models, we only report the specific updates. The updates for the 7 parsimonious structures of $\boldsymbol{\Psi}_g$ are reported in Appendix B.

2.3.5 CM-Step 4

At the fourth CM-step, we first define the "partial" complete-data log-likelihood function

$$\ell_{pc}\left(\boldsymbol{\Omega}; \mathbf{S}_{pc}\right) = \ell_{1c}\left(\boldsymbol{\pi}; \mathbf{S}_{pc}\right) + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig} \ln f_{\text{MVTIN}}(\mathbf{X}_i; \boldsymbol{\Theta}_g), \tag{5}$$

where "partial" refers to fact that the complete data are now defined as $\mathbf{S}_{pc} = \{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^{N}$. Then, keeping fixed $\boldsymbol{\Omega}_1$ at $\ddot{\boldsymbol{\Omega}}_1$ and $\boldsymbol{\Omega}_2$ at $\ddot{\boldsymbol{\Omega}}_2$, we update the tailedness parameters. Notice that, regardless of whether the tailedness parameters are tied or not across the mixture components, we always work with a univariate constrained maximization problem. Specifically, if the tailedness parameters are unconstrained, then $\ddot{\theta}_g$ is determined by maximizing

$$\sum_{i=1}^{N} \ddot{z}_{ig} \ln f_{\text{MVTIN}}(\mathbf{X}_i; \ddot{\boldsymbol{\Theta}}_g)$$

over $\theta_g \in (0, 1)$, $g = 1, \ldots, G$. On the contrary, if $\theta_1 = \cdots = \theta_G = \theta$, then we have to maximize

$$\sum_{i=1}^{N} \sum_{g=1}^{G} \ddot{z}_{ig} \ln f_{\text{MVTIN}}(\mathbf{X}_i; \ddot{\boldsymbol{\Theta}}_g)$$

over $\theta \in (0, 1)$.

2.4 A Note on the Initialization Strategy

Since the likelihood function of finite mixture models is usually multimodal, there exists the possibility for EM-based algorithms to converge to one of the multiple local maximums. This issue is heavily dependent on the choice of the initial values used to start the algorithms [12, 16]. Additionally, bad initial values can lead EM-based algorithms to converge to spurious solutions. As discussed in [11], spurious clusters are made of very few observations, reflect a random pattern in the data rather than a true underlying group structure and are of little practical interest.

With the attempt of addressing these problems, we implemented the EM-EM initialization strategy proposed by [1] and recently used in a matrix-variate framework in [19]. This procedure consists in H short runs of the algorithm from different random positions. The term "short" means that the algorithm is run for a small number of iterations s, without waiting for its convergence. Herein, we set $H = 100$ and $s = 1$. Then, the parameter set producing the largest (observed) log-likelihood is used to initialize the AECM algorithm. At convergence of the algorithm, with the aim of further mitigate the spurious clusters problem, we discarded the solutions having $\pi_g < 0.05$, as done in [9, 21]. In both simulated and real data analyses this procedure has proven stable results.

3 Numerical Analyses

In this section, we conduct a simulation study to assess the ability of the proposed AECM to generate valid estimates of the parameters. Different scenarios are considered. Furthermore, we analyze the ability of the BIC to select the parsimonious structure of the data generating model (DGM). We adopt the original BIC formulation where the lower is its value, the better is the fitting.

Finally, we apply our parsimonious models to a real dataset, along with parsimonious MVN-Ms, to evaluate the fitting performance of both families of models. The detected data partition of the overall best fitting model, as well as the corresponding

estimated parameters, are also commented. Operationally, the functions used to fit our and the competing models are available at the following repository https://github.com/danieletomarchio/Parsimonious-MVTIN-mixtures.

3.1 Simulated Data

3.1.1 Overview

Considering the high number of parsimonious models herein introduced, we only consider the VVE-VE-V MVTIN-Ms as DGM. We consider four different scenarios, that differ by sample size ($N = 200$ and $N = 500$) and by level of separation between groups (O_1 for closer groups and O_2 for more distant groups). In detail, we have

1. Scenario A: $N = 200$ and O_1;
2. Scenario B: $N = 200$ and O_2;
3. Scenario C: $N = 500$ and O_1;
4. Scenario D: $N = 500$ and O_2.

For each scenario, we generated 50 datasets from our DGM with $G = 2$. The parameters used to generate the data, that are common among all the scenarios, are

$$\mathbf{M}_1 = \begin{pmatrix} 0.50 & -0.70 & 0.90 & -0.80 \\ -1.10 & -1.70 & -1.60 & -1.00 \\ 1.40 & 1.50 & 1.30 & 1.20 \end{pmatrix}, \quad \theta_1 = 0.80, \quad \theta_2 = 0.95, \quad \pi_1 = \pi_2 = 0.50$$

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 2.707 & -1.499 & -0.927 \\ -1.499 & 2.738 & 0.983 \\ -0.927 & 0.983 & 1.031 \end{pmatrix}, \quad \mathbf{\Sigma}_2 = \begin{pmatrix} 0.898 & -0.525 & -0.336 \\ -0.525 & 0.908 & 0.359 \\ -0.336 & 0.359 & 0.268 \end{pmatrix},$$

$$\mathbf{\Psi}_1 = \begin{pmatrix} 1.065 & -0.980 & 0.420 & 0.407 \\ -0.980 & 1.812 & -0.858 & -0.156 \\ 0.420 & -0.858 & 1.143 & -0.334 \\ 0.407 & -0.156 & -0.334 & 1.813 \end{pmatrix}, \mathbf{\Psi}_2 = \begin{pmatrix} 1.723 & -2.112 & 1.002 & 0.960 \\ -2.112 & 3.385 & -1.938 & -0.352 \\ 1.002 & -1.938 & 1.785 & -0.874 \\ 0.960 & -0.352 & -0.874 & 3.524 \end{pmatrix}.$$

To obtain $\mathbf{M}_2$, we added a constant c to each element of $\mathbf{M}_1$, where c differs according to whether O_1 or O_2 is considered. In this way, we can control in some way the level of separation among the groups by simply shifting the mean matrices. We set $c = 0.5$ in the scenarios having O_1, and $c = 5$ in the scenarios where O_2 is involved.

Notice that, to simplify the visualization of the parameter recovery results, we follow an approach similar to that used by [6], i.e. we calculate the average across the MSEs of the elements of each estimated parameter over the 2 groups, allowing us to summarize the MSE of each estimated parameter in a single number. Lastly, we take into account the well-known label switching issue by simply attributing the labels according to the values of the estimated mean matrices.

Table 2 Average MSEs of the parameter estimates for the VVE-VE-V MVTIN-Ms. The average is computed among the MSEs of the elements of each estimated parameter, over the $G = 2$ groups and 50 datasets for each scenario

Parameter	Scenario A	Scenario B	Scenario C	Scenario D
π	0.0012	0.0011	0.0005	0.0004
$\mathbf{M}$	0.0568	0.0479	0.0232	0.0199
$\mathbf{\Sigma}$	0.0342	0.0154	0.0335	0.0089
$\mathbf{\Psi}$	0.0332	0.0205	0.0217	0.0075
θ	0.0018	0.0014	0.0011	0.0005

Table 3 Number of times the correct parsimonious structure is detected by the BIC over the 50 datasets in each scenario

Scenario A	Scenario B	Scenario C	Scenario D
36	41	49	50

3.1.2 Results

As concerns the parameter recovery aspect, we fit the VVE-VE-V MVTIN-Ms with $G = 2$ to each generated dataset, and the results are reported in Table 2. As we can see, the average MSEs take quite low values under all the considered scenarios. Furthermore, it is interesting to note that, when we move from Scenario A to Scenario C and from Scenario B to Scenario D (i.e., we increase N for a fixed level of separation between the groups), the estimators become more precise. Additionally, the average MSEs improve as we pass from Scenario A to Scenario B and from Scenario C to Scenario D (i.e., we increase the level of separation between the groups for a fixed N).

Concerning the ability of the BIC in detecting the true parsimonious structure of the DGM, we report in Table 3 the obtained results. We can easily see that the BIC is generally able to detect the true parsimonious structure in the data. Additionally, as we move from Scenario A to Scenario C and from Scenario B to Scenario D, the BIC greatly increases its performance. Also, regardless of the existing separation among the groups, in scenarios C and D the BIC selects the correct parsimonious structures with a probability close to one. The separation among the groups seems to be more relevant when comparing scenarios A and B, with better results obtained in Scenario B, as it might be reasonable to expect. Moreover, we note that in Scenario A the failure to identify the true parsimonious structure is related to the structure of θ_g. The same happens in Scenario B, where 8 times out of 9 the wrong selection is dependent on the chosen parsimonious structure for the θ_g. However, it must be recalled that the tailedness parameters are the most difficult to estimate [19].

Table 4 Parsimonious structure, number of groups G and value of the BIC for the best fitting models belonging to each family

Family	Parsimonious structure	G	BIC
MVN-Ms	VEV-EE	4	59.70
MVTIN-Ms	EEE-EE-V	3	17.09

3.2 Real Data

Here, we analyze the `IMDb` dataset contained in the **MatTransMix** R package [24]. It consists of average ratings for a set of $N = 105$ comedy movies released between 2014 and 2016. For each movie, we have the average ratings divided by gender and age class. Specifically, we have $p = 2$ genders (female and male) and $r = 4$ age classes (0–17, 18–29, 30–44, 45+). Thus, each statistical unit comes in the form of a 2×4 matrix. This dataset has been previously analyzed by [17].

We fitted to this dataset our parsimonious MVTIN-Ms as well as the parsimonious MVN-Ms for $G \in \{1, 2, 3, 4, 5\}$, and the fitting results provided by the best fitting model for each family of models are reported in Table 4. As we can see, the best fitting model among MVN-Ms has $G = 4$ components and a VEV-EE parsimonious structure, whereas the best among MVTIN-Ms has $G = 3$ components and an EEE-EE-V parsimonious structure. However, our model generates a lower BIC. Furthermore, the fitting results for our EEE-EE-V MVTIN-Ms are better than the best fit presented by [17].

By focusing on the EEE-EE-V MVTIN-Ms, Fig. 1 illustrates the parallel coordinate plots of the detected data partition. The three detected groups appear to represent three typologies of movie ratings: low, medium and high, respectively. These indications are also summarized in the estimated mean matrices, that are

$$\mathbf{M}_1 = \begin{pmatrix} 6.37 & 5.66 & 5.45 & 5.47 \\ 5.29 & 5.20 & 5.01 & 5.03 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 7.22 & 6.59 & 6.35 & 6.34 \\ 6.85 & 6.44 & 6.16 & 6.10 \end{pmatrix},$$

$$\mathbf{M}_3 = \begin{pmatrix} 8.03 & 7.68 & 7.51 & 7.53 \\ 7.92 & 7.61 & 7.33 & 7.28 \end{pmatrix}.$$

On average, females assign higher movie ratings than males, and for both genders these ratings generally decrease as the age class increases. The estimated scale matrices are

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.03 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}, \quad \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_3 = \begin{pmatrix} 4.38 & 3.35 & 2.93 & 2.68 \\ 3.35 & 3.44 & 3.18 & 2.92 \\ 2.93 & 3.18 & 3.30 & 3.20 \\ 2.68 & 2.92 & 3.20 & 3.95 \end{pmatrix}.$$

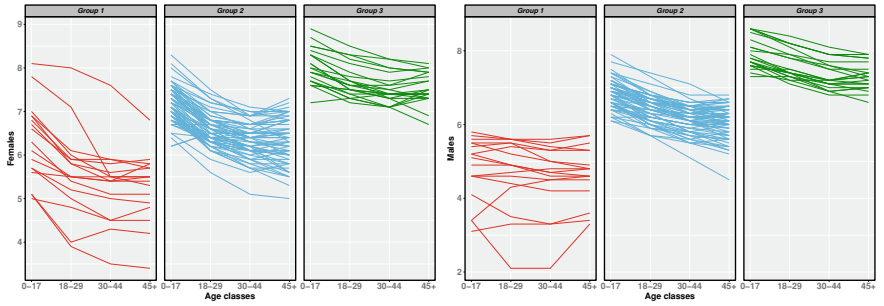


Fig. 1 Parallel coordinate plots constructed for the best clustering solution obtained for the `IMDb` dataset. Left and right plots represent ratings provided by female and male customers, respectively. In turn, each subplot reflect cluster assignments

Thus, we notice similar variabilities among the two genders and a decrease in the variabilities as we move along the first three age classes. Additionally, the EEE-EE scale structure highlights the benefit of considering parsimony with respect to a fully unconstrained model. Lastly, the estimated tailedness parameters are $\theta_1 = 0.99$, $\theta_2 = 0.76$ and $\theta_3 = 0.39$. Therefore, the three clusters are also ordered in terms of heavy-tailedness, with the first group having very heavy tails and the third group having almost normal-tails.

4 Conclusions

Matrix-variate mixture models are a powerful device for clustering matrix-valued data. However, such models tend to be overparameterized primarily because of the large number of parameters involved in the component scale matrices. Furthermore, the tails of the matrix-variate normal distribution, often considered for the functional form of the mixture components, might not be heavy enough to adequately model real data. To jointly address both concerns, in this paper we have introduced a family of 196 parsimonious mixture models, based on the matrix-variate tail-inflated normal distribution, a heavy-tailed generalization of the matrix-variate normal distribution. These models are obtained by imposing constraints on the component scale matrices via the well-known eigen-decomposition, as well as by allowing the tailedness parameters to be tied across the groups. An AECM algorithm for parameter estimation has been presented, and its performance in terms of parameter recovery has been analyzed via simulated data. Furthermore, the ability of the BIC in detecting the parsimonious structure of the true data generating model has been investigated. Under both points of view, the results show an accurate estimation. Finally, our family of models has been fitted along with parsimonious mixtures of matrix-variate normal distributions to a real dataset. Results demonstrate the usefulness of the developed methodology.

Appendix A

Let $\ddot{\mathbf{V}} = \sum_{g=1}^{G} \ddot{\mathbf{V}}_g$, where $\ddot{\mathbf{V}}_g = \sum_{i=1}^{N} \ddot{z}_{ig} \ddot{w}_{ig} \left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right) \dot{\boldsymbol{\Psi}}_g^{-1} \left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)'$. Then, we have the following updates:

- Model EII

$$\ddot{\lambda} = \frac{\operatorname{tr}\left\{\ddot{\mathbf{V}}\right\}}{prN};$$

- Model VII

$$\ddot{\lambda}_g = \frac{\operatorname{tr}\left\{\ddot{\mathbf{V}}_g\right\}}{pr\sum_{i=1}^{N} \ddot{z}_{ig}};$$

- Model EEI

$$\ddot{\boldsymbol{\Delta}} = \frac{\operatorname{diag}\left(\ddot{\mathbf{V}}\right)}{\left|\operatorname{diag}\left(\ddot{\mathbf{V}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\left|\operatorname{diag}\left(\ddot{\mathbf{V}}\right)\right|^{\frac{1}{p}}}{rN};$$

- Model VEI

$$\ddot{\boldsymbol{\Delta}} = \frac{\operatorname{diag}\left(\sum_{g=1}^{G} \dot{\lambda}_g^{-1} \ddot{\mathbf{V}}_g\right)}{\left|\operatorname{diag}\left(\sum_{g=1}^{G} \dot{\lambda}_g^{-1} \ddot{\mathbf{V}}_g\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_g = \frac{\operatorname{tr}\left\{\ddot{\boldsymbol{\Delta}}^{-1} \ddot{\mathbf{V}}_g\right\}}{pr\sum_{i=1}^{N} \ddot{z}_{ig}};$$

- Model EVI

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)}{\left|\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{g=1}^{G} \left|\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)\right|^{\frac{1}{p}}}{rN};$$

- Model VVI

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)}{\left|\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_g = \frac{\left|\operatorname{diag}\left(\ddot{\mathbf{V}}_g\right)\right|^{\frac{1}{p}}}{r\sum_{i=1}^{N} \ddot{z}_{ig}};$$

- Model EEE

$$\ddot{\boldsymbol{\Sigma}} = \frac{\ddot{\mathbf{V}}}{rN};$$

- Model VEE

$$\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}\ddot{\boldsymbol{\Gamma}}' = \frac{\sum_{g=1}^{G}\dot{\lambda}_g^{-1}\ddot{\mathbf{V}}_g}{\left|\sum_{g=1}^{G}\dot{\lambda}_g^{-1}\ddot{\mathbf{V}}_g\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_g = \frac{\text{tr}\left\{(\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}\ddot{\boldsymbol{\Gamma}}')^{-1}\ddot{\mathbf{V}}_g\right\}}{pr\sum_{i=1}^{N}\ddot{z}_{ig}};$$

- Model EVE
  For this model, there is no analytical solution for $\boldsymbol{\Gamma}$. Thus, an iterative minorization-maximization (MM) algorithm [2] is implemented. Specifically, the following surrogate function is defined

$$f(\boldsymbol{\Gamma}) = \sum_{g=1}^{G}\text{tr}\left\{\mathbf{V}_g\boldsymbol{\Gamma}\boldsymbol{\Delta}_k^{-1}\boldsymbol{\Gamma}'\right\} \leq S + \text{tr}\left\{\dot{\mathbf{F}}\boldsymbol{\Gamma}\right\},$$

  where S is a constant and $\dot{\mathbf{F}} = \sum_{g=1}^{G}\left(\boldsymbol{\Delta}_k^{-1}\dot{\boldsymbol{\Gamma}}'\mathbf{V}_g - e_g\boldsymbol{\Delta}_k^{-1}\dot{\boldsymbol{\Gamma}}'\right)$, with e_g being the largest eigenvalue of $\mathbf{V}_g$. The update of $\boldsymbol{\Gamma}$ is given by $\ddot{\boldsymbol{\Gamma}} = \dot{\mathbf{G}}\dot{\mathbf{H}}'$, where $\dot{\mathbf{G}}$ and $\dot{\mathbf{H}}$ are obtained from the singular value decomposition of $\dot{\mathbf{F}}$. This process is repeated until a specified convergence criterion is met and the estimate $\ddot{\boldsymbol{\Gamma}}$ is obtained from the last iteration. Then, we obtain

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\text{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\ddot{\boldsymbol{\Gamma}}\right)}{\left|\text{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{g=1}^{G}\text{tr}\left(\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}_g^{-1}\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\right)}{prN};$$

- Model VVE
  Similarly to the EVE case, there is no analytical solution for $\boldsymbol{\Gamma}$, and the MM algorithm described above is implemented. Then, we have

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\text{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\ddot{\boldsymbol{\Gamma}}\right)}{\left|\text{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_g = \frac{\left|\text{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{V}}_g\ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}}{r\sum_{i=1}^{N}\ddot{z}_{ig}};$$

- Model EEV
  Consider the eigen-decomposition $\mathbf{V}_g = \mathbf{L}_g\mathbf{D}_g\mathbf{L}_g'$, with eigenvalues in the diagonal matrix $\mathbf{D}_g$ following descending order and orthogonal matrix $\mathbf{L}_g$ composed of the corresponding eigenvectors. Then, we obtain

$$\ddot{\boldsymbol{\Gamma}}_g = \ddot{\mathbf{L}}_g, \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{g=1}^{G} \ddot{\mathbf{D}}_g}{\left|\sum_{g=1}^{G} \ddot{\mathbf{D}}_g\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\left|\sum_{g=1}^{G} \ddot{\mathbf{D}}_g\right|^{\frac{1}{p}}}{rN};$$

- Model VEV
  By using the same algorithm applied for the EEV model, we have

$$\ddot{\boldsymbol{\Gamma}}_g = \ddot{\mathbf{L}}_g, \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{g=1}^{G} \lambda_g^{-1} \ddot{\mathbf{D}}_g}{\left|\sum_{g=1}^{G} \lambda_g^{-1} \ddot{\mathbf{D}}_g\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_g = \frac{\operatorname{tr}\left\{\ddot{\mathbf{D}}_g \ddot{\boldsymbol{\Delta}}^{-1}\right\}}{pr \sum_{i=1}^{N} \ddot{z}_{ig}};$$

- Model EVV

$$\ddot{\boldsymbol{\Gamma}}_g \ddot{\boldsymbol{\Delta}}_g \ddot{\boldsymbol{\Gamma}}_g' = \frac{\ddot{\mathbf{V}}_g}{\left|\ddot{\mathbf{V}}_g\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{g=1}^{G} \left|\ddot{\mathbf{V}}_g\right|^{\frac{1}{p}}}{rN};$$

- Model VVV

$$\ddot{\boldsymbol{\Sigma}}_g = \frac{\ddot{\mathbf{V}}_g}{r \sum_{i=1}^{N} \ddot{z}_{ig}}.$$

Appendix B

Let $\ddot{\mathbf{W}} = \sum_{g=1}^{G} \ddot{\mathbf{W}}_g$, where $\ddot{\mathbf{W}}_g = \sum_{i=1}^{N} \ddot{z}_{ig} \ddot{w}_{ig} \left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)' \ddot{\boldsymbol{\Sigma}}_g^{-1} \left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)$. With the exclusion of the II model, for which no parameters need to be estimated, we have the following updates:

- Model EI

$$\ddot{\boldsymbol{\Delta}} = \frac{\operatorname{diag}\left(\ddot{\mathbf{W}}\right)}{\left|\operatorname{diag}\left(\ddot{\mathbf{W}}\right)\right|^{\frac{1}{r}}};$$

- Model VI

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\operatorname{diag}\left(\ddot{\mathbf{W}}_g\right)}{\left|\operatorname{diag}\left(\ddot{\mathbf{W}}_g\right)\right|^{\frac{1}{r}}};$$

- Model EE

$$\ddot{\boldsymbol{\Psi}} = \frac{\ddot{\mathbf{W}}}{\left|\ddot{\mathbf{W}}\right|^{\frac{1}{r}}};$$

- Model VE
  As for the EVE and VVE models, there is no analytical solution for $\boldsymbol{\Gamma}$ and a MM algorithm of the type described for the EVE model is implemented, after replacing $\mathbf{V}$ with $\mathbf{W}$. Then, we have

$$\ddot{\boldsymbol{\Delta}}_g = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{W}}_g\ddot{\boldsymbol{\Gamma}}\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}'\ddot{\mathbf{W}}_g\ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{r}}};$$

- Model EV
  By using the same approach of the EEV and VEV models, after replacing $\ddot{\mathbf{V}}$ with $\ddot{\mathbf{W}}$, we have

$$\ddot{\boldsymbol{\Gamma}}_g = \ddot{\mathbf{L}}_g \quad \text{and} \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{g=1}^{G}\ddot{\mathbf{D}}_g}{\left|\sum_{g=1}^{G}\ddot{\mathbf{D}}_g\right|^{\frac{1}{r}}};$$

- Model VV

$$\ddot{\boldsymbol{\Psi}}_g = \frac{\ddot{\mathbf{W}}_g}{\left|\ddot{\mathbf{W}}_g\right|^{\frac{1}{r}}}.$$

References

1. Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Stat. Data Anal. **41**(3–4), 561–575 (2003)
2. Browne, R.P., McNicholas, P.D.: Estimating common principal components in high dimensions. Adv. Data Anal. Classific. **8**(2), 217–226 (2014)
3. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**(5), 781–793 (1995)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–22 (1977)

5. Doğru, F.Z., Bulut, Y.M., Arslan, O.: Finite mixtures of matrix variate t distributions. Gazi Univ. J. Sci. **29**(2), 335–341 (2016)
6. Farcomeni, A., Punzo, A.: Robust model-based clustering with mild and gross outliers. Test **29**(4), 989–1007 (2020)
7. Gallaugher, M.P.B., McNicholas, P.D.: Finite mixtures of skewed matrix variate distributions. Pattern Recognit. **80**, 83–93 (2018)
8. Gupta, A.K., Varga, T., Bodnar, T.: Elliptically Contoured Models in Statistics and Portfolio Theory. Springer, New York (2013)
9. Leisch, F.: Flexmix: a general framework for finite mixture models and latent glass regression in R. J. Stat. Softw. **11**(8), 1–18 (2004)
10. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley (2007)
11. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
12. Melnykov, V., Melnykov, I.: Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. Comput. Stat. Data Anal. **56**(6), 1381–1395 (2012)
13. Melnykov, V., Zhu, X.: On model-based clustering of skewed matrix data. J. Multivar. Anal. **167**, 181–194 (2018)
14. Melnykov, V., Zhu, X.: Studying crime trends in the USA over the years 2000–2012. Adv. Data Anal. Classific. **13**(1), 325–341 (2019)
15. Meng, X.L., Van Dyk, D.: The EM algorithm-an old folk-song sung to a fast new tune. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **59**(3), 511–567 (1997)
16. Michael, S., Melnykov, V.: An effective strategy for initializing the EM algorithm in finite mixture models. Adv. Data Anal. Classific. **10**(4), 563–583 (2016)
17. Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)
18. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
19. Tomarchio, S.D., Punzo, A., Bagnato, L.: Two new matrix-variate distributions with application in model-based clustering. Comput. Stat. Data Anal. **152**, 107050 (2020)
20. Tomarchio, S.D., Gallaugher, M.P.B., Punzo, A., McNicholas, P.D.: Mixtures of matrix-variate contaminated normal distributions. J. Comput. Graph. Stat. **31**(2), 413–421 (2022)
21. Tomarchio, S.D., McNicholas, P.D., Punzo, A.: Matrix normal cluster-weighted models. J. Classific. **38**(3), 556–575 (2021)
22. Viroli, C.: Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. **21**(4), 511–522 (2011)
23. Viroli, C.: Model based clustering for three-way data structures. Bayesian Anal. **6**(4), 573–602 (2011)
24. Zhu, X., Melnykov V.: MatTransMix: an R package for clustering matrices. R package version 0.1.15 (2021)

Population Size Estimation by Repeated Identifications of Units. A Bayesian Semi-parametric Mixture Model Approach

Tiziana Tuoto, Davide Di Cecco, and Andrea Tancredi

Abstract The use of mixture models for estimating the size of an elusive population when capture rates vary among individuals has received strong attention from researchers. In this paper we propose a Bayesian semi-parametric approach by considering a truncated infinite dimensional Poisson mixture model for capture recapture count data. An application in official statistics regarding the estimate of the size of criminal populations is used to illustrate the proposed methodology.

Keywords Criminal populations · Capture-recapture · Dirichlet process mixture · Official statistics

1 Introduction

The aim of this paper is to estimate the size of a hidden criminal population, for instance people working in markets of drug trafficking, prostitution exploitation and smuggling in Italy during a given reference time. The estimate of the size of people involved in these kinds of illegal economic activities, is envisaged at European level: according to European regulations, national accounts aggregates have to include illegal activities covering exhaustively the economic transactions which occur in the economic system.

In this paper, we aim at estimating the size of people involved in smuggling of goods. In Italy, smuggling activities mainly regards cigarettes, and they are related to the importation and exportation of products that are legal in some other countries. Illegal cigarettes arrive in Italy especially from Eastern European countries, China

T. Tuoto (✉)
ISTAT, Rome, Italy
e-mail: tuoto@istat.it

D. Di Cecco · A. Tancredi
Università di Roma La Sapienza, Rome, Italy
e-mail: davide.dicecco@uniroma1.it

A. Tancredi
e-mail: andrea.tancredi@uniroma1.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_25

and the United Arab Emirates. It is worthwhile nothing that there are other economic aspects, such as organized crime and corruption of the legal economy by money laundering that underpin these activities to facilitate them. The need to estimate the size of illegal populations is driven also by social and judicial factors, it helps to better understand the size of the illegal phenomenon itself, and to assess the threat it poses to society; to better size the police forces to counter it and to evaluate the effectiveness of prevention and counteraction policies.

Illegal activities for their nature are difficult to measure as people involved have obvious reasons to hide these activities. In this study we exploit administrative registers coming from the Ministry of Justice, which report alleged crimes for which the judicial authority started a criminal proceeding. Crimes records in the registers of the Public Prosecutor's offices, contain soft identifiers of the denounced subjects, namely date and place of birth and gender.

On the basis of the soft identifiers, crime authors can be identified and followed in a specific time span. In this way, the administrative source can be considered as a list of potential criminals with the count (i.e. the number of times) that they appear in the Prosecutor's offices registers. In the list we can observe individuals who are charged $1, 2, 3, \ldots,$ times, however we cannot observe units not caught by the Justice system. Hence, the registers can be considered as incomplete lists of potential criminals, since only denounced crimes and suspected criminals are reported. We want to estimate the hidden part of the population, i.e. the size of it not reported on the registers of the Public Prosecutor's offices. We assume that the population of potential criminals in a given year is a closed population of unknown size. This is a classic assumption that rarely fully hold in human populations, since we cannot exclude that the population does not change, no entries and/or exits, during a year. To verify how realistic this assumption is for the population at hand would require additional information on the captures time which are not available to us.

Notice that the capture recapture data for the problem at hand are usually called repeated counting data. The common parametric approach to analyse this data is to define a counting distribution for the number of captures in the population, and the use of mixture distributions to model individual heterogeneity in the captures probabilities represents a standard approach, in the frequentist literature, see for example [1–3]. Here we follow a Bayesian point of view founding the analysis on the Dirichlet process mixture model.

The paper is organised as follows: in the next Section we specify the model and outline the resulting simulation algorithm. In Sect. 3 we compare a Dirichlet process mixture model (DPM) and a sparse finite mixture (SFM). SFM represent an interesting alternative to the DPM, in particular when the number of components does not increase with the size of the observed sample, as it seems reasonable in capture-recapture context. However, to the best of our knowledge, this class of models has not yet been applied in this field. Finally, in the last Section we briefly illustrate the resulting posterior inference on the size of the smugglers in 2014 in Italy, comparing it with other Bayesian and frequentist methods.

2 The Model

We assume that the population of potential criminals in a given year is a closed population of unknown size N. To take into account the heterogeneity in criminals captures we assume that the number of times Y that a criminal appears in the Prosecutor's offices registers is a mixture of Poisson distributions. In particular, we assume the *truncated* version of the Dirichlet process mixture, see [4], where the weights of a fixed number Poisson components follow a finite stick breaking process.

Denote as k the number of components. Moreover denote as $p_j = Prob(Y = j)$ the probability of a unit being captured j times, and as p_j^i the probability of being captured j times in the ith component, $p_j^i = \lambda_i^j e^{-\lambda_i}/j!$. Finally let $\pi_1, \ldots, \pi_k$ be the mixing weights, so that

$$p_j = \sum_{i=1}^{k} \pi_i \; p_j^i \quad j = 0, 1 \ldots . \tag{1}$$

Denote as n_j the number of observed units that have been captured j times, and as D the set of all observed counts, $D = \{n_j\}_{j>0}$. We have $\sum_{j>0} n_j = n_{obs}$. We want to estimate the number of uncaptured units n_0 or, equivalently, the total number of units in the population $N = n_{obs} + n_0$. Note that n_0 is Binomial(N, p_0).
We set a conjugate Gamma prior for each parameter λ_i, a truncated stick–breaking process with parameter ϕ over the mixture weights, and a Gamma prior over ϕ.

$$\begin{aligned} \lambda_i &\sim Gamma(\alpha_i, \beta_i), \quad i = 1, ..., k^* \\ (\pi_1, ..., \pi_{k^*}) &\sim SB(\phi) \\ \phi &\sim Gamma(\alpha_\phi, \beta_\phi) \end{aligned}$$

A similar modeling approach was considered by [5] in the standard multiple system framework with a fixed number of lists. In particular a Dirichlet process mixture was proposed to model the heterogeneity in the capture histories. In the context of repeated count data for modeling gene expression sequence abundance, a semi-parametric mixture of Poisson driven by the Dirichlet process with the censoring of zero counts was proposed by [6].

2.1 MCMC Algorithm

In this Section we detail the Gibbs–based MCMC algorithm to sample from the posterior distribution of N. Let us denote as Θ all the parameters $\{\pi_i\}$ and $\{\lambda_i\}$. Moreover let n_j^i be the (latent) number of units in the ith component that have been captured j times. Let n^i be the total number of population units (captured or

uncaptured) in component i: $n^i = \sum_{j \geq 0} n^i_j$. Then, at iteration t we have the following steps:

1. Sample all parameters $\lambda_i^{(t)}$

$$\lambda_i \sim Gamma\left(\sum_{j \geq 0} j \cdot n^i_j + \alpha_i \,,\; n^i + \beta_i\right) \qquad \text{for } i = 1, \ldots, k$$

2. In order to sample all mixing weights $\pi_i^{(t)}$, we first sample

$$V_i \sim Beta\left(1 + n^i, \phi + \sum_{h=i+1}^{k} n^h\right) \qquad i = 1, \ldots, k-1$$

then take

$$\pi_i = V_i \prod_{h, h<i} (1 - V_h) \qquad \text{for } i = 1, ..., k$$

where $V_k = 1$.
3. Sample $\phi^{(t)}$, $\phi \sim Gamma(\alpha_\phi - 1 + k \,,\; \beta_\phi - \log \pi_k)$
4. Sample $N^{(t)}$ from $P(N \mid \Theta^{(t)}, D)$. Note that

$$\begin{aligned} P(N \mid \Theta, D) = P(N \mid \Theta, n_{obs}) &\propto P(N) P(n_{obs} \mid N, \Theta) \\ &\propto P(N) \binom{N}{n_{obs}} p_0^{N - n_{obs}} (1 - p_0)^{n_{obs}}, \end{aligned} \tag{2}$$

where the probability p_0 of not being captured is calculated according to (1). Then, if we choose the improper prior $P(N) \propto 1/N$, we have

$$N^{(t)} \sim NegBin\left(n_{obs}, 1 - p_0^{(t)}\right).$$

Other (possibly informative) choices for the prior over N can be easily managed with a Metropolis step.
5. Sample vector $\mathbf{n}_j^{(t)} = (n_j^1, ..., n_j^k)$ from $P(\mathbf{N}_j \mid \Theta^{(t)}, N^{(t)}, D)$:

$$\mathbf{N}_j \sim Mult\left(n_j, (p_{1|j}, ..., p_{k|j})\right) \qquad \text{for } j \geq 0$$

where $n_0 = N^{(t)} - n_{obs}$, and the probabilities $p_{i|j}$ of belonging to the i–th component conditionally on the number of captures j and the current values of Θ are updated as:

$$p_{i|j} = \frac{\pi_i \, p_j^i}{\sum_{h=1}^{k} \pi_h \, p_j^h}.$$

3 Sparse Finite Mixtures

An interesting alternative to the Dirichlet process mixture is represented by the sparse finite mixture, see [7, 8]. Sparse finite mixtures are obtained by assuming a sparse symmetric Dirichlet prior on the component weights of an overfitting finite mixture distribution, that is a distribution where the number of components is larger than the number of clusters in the data. In this way, the number of clusters in the data is not fixed a priori, rather, as for Dirichlet process mixtures, it is random by construction and can be estimated from the data using standard MCMC methods. As shown by [9], Dirichlet process mixture can be seen as the limiting case of a sparse finite mixture.

In our context, a favorable property of sparse finite mixture consists in its behaviour when the number of observations n_{obs} increases, since in this case, differently from DPM, sparse finite mixture avoids to create new cluster, even if n_{obs} goes to infinity. On the other side [10] demonstrate that DPM tends to increase the number of clusters with n_{obs}, that is, it is very likely that one big cluster is found, the sizes of further clusters geometrically decay, and many singleton clusters are estimated. Also [11] discuss the properties of Dirichlet process mixtures, with respect to the number of components, highlighting the risk of overestimating the number of clusters and producing not consistent estimates.

Anyway notice that, as shown by [12], the clustering performance of both Dirichlet processes mixture and sparse finite mixture becomes comparable under specific prior assumptions. In fact, if the prior on the precision parameter of the DPM and the symmetric Dirichlet driving the stick-breaking representation are appropriately chosen and matched, sparse clustering can be obtained by both model classes. In particular, overfitting of the number of clusters for DPM can be avoided by using a prior on the precision parameter ϕ that favors very small values for the mixture components and the consequence sparsity in the mixture distribution. Note also that when using these sparse priors, the posterior distribution of the number of clusters will be more sensitive on the prior hyperparameters with respect to a standard fitting of the DPM or a finte mixture model but the two sparse clustering approach will provide similar results

We now fix the notation for sparse finite mixtures. Assuming a prior Dirichlet distribution on the components weights $\pi = (\pi_1, \ldots, \pi_k)$, i.e. $\pi \sim D(e_1, \ldots, e_k)$. symmetry is obtained by taking $e_i \equiv e_0, i = 1, \ldots, k$; such a prior being denoted by $\pi \sim DK(e_0)$. If e_0 is a small value, then many of the k weights will be small a priori, implying that not all k components will generate a cluster of their own. The size of non-empty clusters depends on e_0 and k. Note that by increasing k and e_0 also the expected number of non-empty clusters increases. Finally note that posterior inference for sparse finite mixture model can be obtained by the standard Gibbs sampler algorithm for finite mixtures [7, 8].

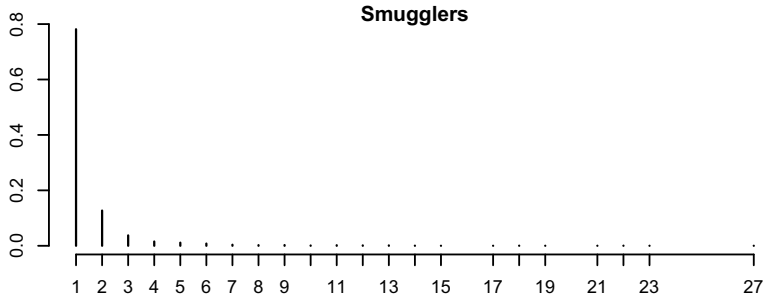


Fig. 1 Relative frequencies of observed counts for smuggling crimes in 2014

4 Results on Smuggling Data

In this Section, we apply our model to estimate the number of people implicated in smuggling of goods in 2014 in Italy. To this purpose, we exploit administrative registers coming from the Ministry of Justice, which report alleged crimes for which the judicial authority started a criminal proceeding. These data are generally utilized by National Statistical Offices to produce official crime statistics. These statistics refer to crimes reported to the police from victims and witnesses. We propose a re-use of these sources to estimate the people involved in criminal activities but unreported to the justice system. The distribution of observed counts for smuggling captures is reported in Fig. 1. We consider a data set with a total number of observed smugglers equal to $n = 3349$. Note also the fat right tail of the capture distribution with a maximum number of captures equal to 27.

We apply the DPM model with the following prior settings: we set $\phi \sim Gamma(1, 20)$ and a hierarchical prior for the component-specific parameters, that is $\lambda_i|\beta_\lambda \sim Gamma(\alpha_\lambda, \beta_\lambda)$ and $\beta_\lambda \sim Gamma(g_0, G_0)$, where $\alpha_\lambda = 0.1$, $g_0 = 0.5$, and $G_0 = g_0\bar{y}/\alpha_\lambda$, with $\bar{y}$ being the observed mean of the data.

The use of a sparsity prior for the precision parameter ϕ allows us to avoid overfitting the number of clusters for DPM. In this way, DPM models produce fully comparable results with sparse finite mixture models. We also verified that different values of k do not affect the number of components we detect, and actually with prior on ϕ favoring sparse mixtures we always identified the same number of components and the parameters estimates are not affected by the choice of k. This clearly emerges for Fig. 2, that shows box–plots of posterior distributions for N obtained by fixing different values for k, namely from 3 to 15. The behaviour depicted in Fig. 2 reassures us that the choice of the truncated number of components from the infinite mixture does not affect the estimate of the quantity of interest, as this is constant for different values of k. In addition, we have verified that the components detected by the DPM and the resulting estimates are the same as obtained with sparse finite mixtures. These results allow us to safely choose a relatively small value for the truncated number of components k, with the two-fold advantage of reducing the

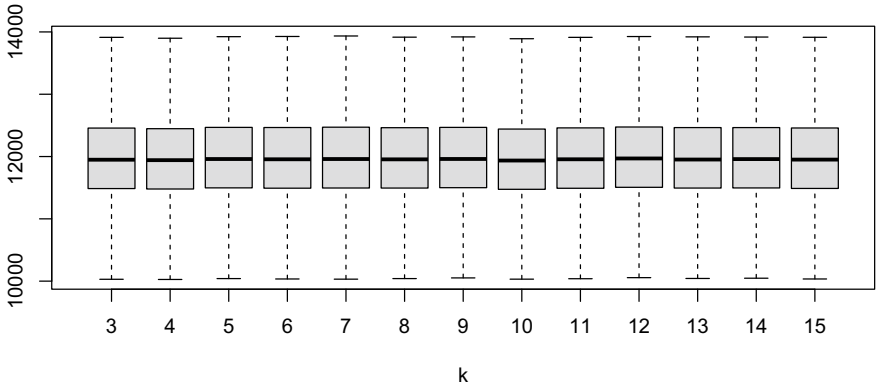


Fig. 2 Posterior distribution of N by truncated DPM with different k values

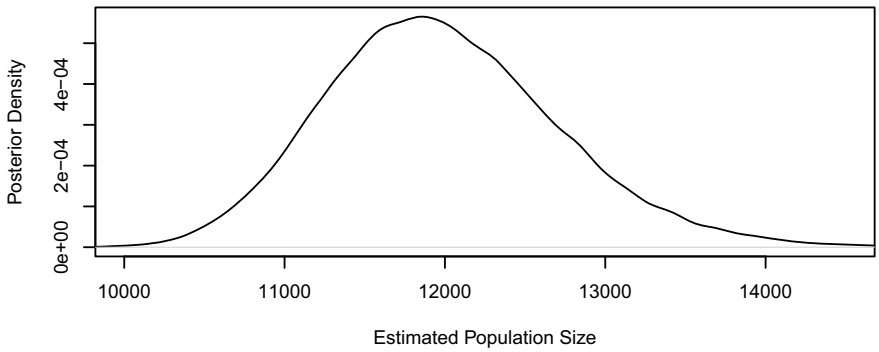


Fig. 3 Posterior density of population size N estimated by DPM model for smuggling crimes in 2014

computational complexity of the DPM algorithm and of avoiding model selection with respect to the number of components as in traditional mixture models.

As far as regards the prior for λ, our setting allows the DPMs to identify the long right tail of the observed y_i values. Indeed, DPM models return analogous results if we provide substantial prior probability to large values, by means of centering the scale parameter of the hyper–prior distribution for λ around 1/20.

The number of replications of the MCMC algorithm is 10^6 with a thinning of 20 observations. Standard diagnostic tools confirmed the convergence of the algorithm. The posterior density for the population size N estimated with $k = 5$ is shown in Fig. 3.

Table 1 summaries results and shows the posterior means and credible intervals of N. For the sake of brevity, Table 1 shows the the posterior mean and the posterior 95% credible interval for N obtained by DPM under the specified setting; estimates from sparse finite mixture model, and from hyper–prior for λ favouring long tails provide very similar figures. In Table 1 we compare the DPM estimates with the ones obtained

Table 1 Population size estimates N (credible intervals), for smuggling data

	Smugglers	
Model	$\hat{N}$	95% C.I.
DPM	12093	10692–13592
Poisson	5583	5392–5774
Negative Binomial	71121	54197–96235
Chao	11387	10451–12447
Zelterman	12052	10952–13152
NPML	12018	10043–13339

assuming a single Poisson distribution and its generalization for heterogeneity, the Negative Binomial distribution. Table 1 shows also some frequentist estimates: the well-known Chao lower bound (see [13] for a review); the Zelterman estimator and the non parametric maximum likelihood (NPML) estimate proposed by [2]. The NPML estimate is calculated by the SPECIES R package, as well as the Chao lower bound. The NPML estimate is obtained from the full likelihood based on a Poisson mixture model, and its 95% confidence interval is obtained by the bootstrap procedure proposed in [14].

Results in Table 1 clearly show the need to take into account the heterogeneity in the estimate of the population size, and how our proposal is effective and appropriate in managing this situation. Indeed, the lower bound Chao estimate providing greater value than the Poisson distribution is a strong evidence of the presence of the heterogeneity in our data. Unfortunately, the Negative Binomial distribution is not a good solution to represent heterogeneity in this case. Actually, it runs into the so-called boundary problem (see, e.g., [15]), that is the Negative Binomial model severely overestimates N, sometimes by several orders of magnitudes, when in the observed (truncated) data the mean number of captures is close to one (in our data we observe $\bar{y} = 1.53$) and therefore the estimate of the size parameter r approaches zero. In Table 1, the Negative Binomial model produces estimates greater than the others models up to six times, much larger credible interval, and the posterior mean estimate for the size parameter r is 0.06. All these factors clearly point out the presence of the so–called boundary problem to such an extent that we are inclined to ignore the estimates coming from this model.

The Zelterman and the non parametric maximum likelihood (NPML) Poisson mixture proposed by [2] produce estimates quite close to the DPM model, and overlapping 95% confidence/credible intervals. The NPML model identifies 3 components, the same as the DPM. This result was quite expected, since we used non-informative priors for all our parameters, confirming in addition the advantage of DPMs which avoid complex and cumbersome procedures for assessing the number of components as in traditional mixture models.

For further comparison, Table 2 shows the estimates for the Poisson mixture components by DPM and NPML models, as well as the estimates of the λ parameter

Table 2 Estimated Poisson mixture components by different models

	Parameter	Component 1	Component 2	Component 3
DPM	λ_i	0.297	3.523	13.055
	π_i	0.970	0.026	0.003
NPML	λ_i	0.298	3.528	13.046
	π_i	0.970	0.026	0.003
Poisson	λ	0.916		
Zelterman	λ	0.163		

for the single Poisson distribution and the Zelterman estimator. Results in this Table further confirm that the Poisson model is not good enough to represent heterogeneity in the data, and its estimate for the parameter λ results too much concentrated on one. The Zelterman estimator is robust with respect to heterogeneity, while both the Bayesian and frequentist Poisson mixtures allow one to identify three components, with very similar λ parameters and relative weights.

5 Concluding Remarks

In this work we presented a Bayesian approach to the analysis of heterogeneity in capture–recapture, based on Dirichlet Process Mixture models.

We showed in an application to real data that this class of model is flexible enough to represent unobserved heterogeneity, and it represents an easy and consistent alternative to other methodologies for choosing the number of components in a Poisson mixture model. The choice of non-informative prior on the DPM precision parameter minimizes the risk of overestimating k, without precluding the use of eventual prior information over the other parameters (typically over N). We also want to stress the incredible computational advantage coming from the use of a *truncated* DPM. In fact, our algorithm runs tens of millions of iterations in less than a minute (compare with [6]).

In this application the NPML approach of Norris and Pollock (1998) and our model produce very similar results. However with other data sets we have observed different estimates with the NPML approach having problems in reporting reliable confidence intervals. Our approach results extremely advantageous also for the assessment of the credible intervals, compared to the bootstrap procedure utilized by the NPML. Indeed, by using different data sets we experienced the NPML is not able to provide a confidence interval in a reasonable time-span, particularly when the size of observed sample and the number of components increase, whilst our algorithm has high performance and produces stable results.

Natural extensions of this work are related to the introduction of individual covariates to cope with observed and unobserved heterogeneity at the same time.

The inclusion of individual covariates in DPM models for capture–recapture analysis seems an unexplored and intriguing topic which deserve future research.

Finally we want to remark that our approach permits to account for the uncertainty in the number of components. In the application presented in this paper Bayesian inference produces very similar results to the frequentist approaches which require a fixed number of components. Anyway, by using different data sets we experienced contrasting results between the two inferential methods, hence it would be important to conduct a deeper comparison between the two inferential strategies. A simulation study would be helpful to compare our proposal and the NPML, also in the light of the well-known identification issues in capture–recapture models.

References

1. Böhning, D., Dietz, E., Kuhnert, R., Schön, D.: Mixture models for capture-recapture count data. Stat. Methods Appl. **14**, 29–43 (2005)
2. Norris, J.L., Pollock, K.H.: Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics **52**, 639–649 (1996)
3. Pledger, S., Pollock, K.H., Norris, J.L.: Open capture-recapture models with heterogeneity: I. Cormack-Jolly-Seber model. Biometrics **59**, 786–794 (2003)
4. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)
5. Manrique-Vallier, D.: Bayesian population size estimation using Dirichlet Process mixtures. Biometrics **72**, 1246–1254 (2016)
6. Guindani, M., Sepulveda, N., Paulino, C.D., Muller, P.: A Bayesian semiparametric approach for the differential analysis of sequence counts data. J. Roy. Stat. Soc. Ser. C **63**, 385–405 (2014)
7. Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Model-based clustering based on sparse finite Gaussian mixtures. Stat. Comput. **26**, 303–324 (2016)
8. Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Identifying mixtures of mixtures using Bayesian estimation. J. Comput. Graph. Stat. **26**, 285–295 (2017)
9. Green, P.J., Richardson, S.: Modelling heterogeneity with and without the Dirichlet process. Scand. J. Stat. **28**, 355–375 (2001)
10. Müller, P., Mitra, R.: Bayesian nonparametric inference—why and how. Bayesian Anal. (Online), **8** (2013)
11. Miller, J.W., Harrison, M.T.: A simple example of Dirichlet process mixture inconsistency for the number of components. Adv. Neural Inf. Process. Syst. 199–206 (2013)
12. Frühwirth-Schnatter, S., Malsiner-Walli, G.: From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. Adv. Data Anal. Classif. **13**, 33–64 (2019)
13. Chao, A.: Capture-Recapture for Human Populations, pp. 1–16. Statistics Reference Online, Wiley StatsRef (2014)
14. Böhning, D., Schön, D.: Nonparametric maximum likelihood estimation of population size based on the counting distribution. J. Roy. Stat. Soc: Ser. C (Appl. Stat.) **54**(4), 721–737 (2005)
15. Böhning, D.: Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. Metron **73**, 201–216 (2015)

Spatial Interdependence of Mycorrhizal Nuclear Size in Confocal Microscopy

Ivan Sciascia, Andrea Crosino, Gennaro Carotenuto, and Andrea Genre

Abstract Our work analyzes the spatial distribution of plant roots cellular nuclei inoculated with Arbuscular mycorrhizal (AM) fungi, and observed on confocal microscope. The images obtained are processed to measure the distribution of the nuclei on the x, y plane and to evaluate the autocorrelation of the nuclear sections size. The image processing protocol proposed involves the density calculation, the Clarks and Evans (Ecology 35:445–453, 1954) index for the positioning and the variogram analysis of the correlation between nuclei size.

Keywords Spatial analysis · Quantitative microscopy · Autocorrelation · Nearest neighbor index · Variogram · Kriging

1 Introduction

In various geostatistical studies, analysis techniques aimed at the spatial distribution of objects have been introduced. In forest sciences it is possible to analyze a stand and the characteristics of the trees in terms of density, spatial point distribution, mark point process analysis, considering variations in stand characteristics such as basal area, diameter of trees, for estimating biomass production [5].

Geostatistics and the study of regionalized variables developed by Matheron [6] has been applied in several fields including forestry, agriculture, ecology to describe spatial variations in the characteristics of natural resources. Spatial statistics and geostatistics have been used originally to describe the variation of natural

I. Sciascia (✉) · A. Crosino · G. Carotenuto · A. Genre
Department of Life Sciences and Systems Biology, Viale Mattioli 25, Torino, Italy
e-mail: ivan.sciascia@unito.it

A. Crosino
e-mail: andrea.crosino@unito.it

G. Carotenuto
e-mail: gennaro.carotenuto@unito.it

A. Genre
e-mail: andrea.genre@unito.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_26

phenomena, such as the distribution of mineral resources and successively adopted in other research fields [3]. Mathematical models that fit functions to interpolate experimental spatial point distribution of natural features have been described in several papers [3, 4].

Spatial statistics analysis has also been conducted to predict continuous variables in forest science using global forest inventory data to make estimates for small areas. The forest variables involved for the estimate are for example the average diameter, average height, average age, basal area and volume, which in any case may not be spatially continuous as they can strongly depend on the structure of the forest landscape, its boundaries and commercial management [5].

In our work, the spatial statistical analysis techniques are transported in confocal microscopy image analysis developing tools for the quantitative microscopy.

These techniques could have good success for the estimation of biophysical parameters of cell morphology.

In the present study we consider visual evaluation of morphology coupled with computing patterns of image recognition. The goals are to calculate a sampling of fluorescent market nuclei sections in root cells inoculated or not with AM fungi and to estimate point pattern spatial distribution and autocorrelation of nuclei sections size.

The studies in quantitative microscopy consider visual evaluation of morphology coupled with computing patterns of image recognition. They use optical microscopy and biotechnology techniques to mark objects using fluorescent probes; it allows to calculate and estimate cellular parameters, such as distance between nuclei and size [1] in regions of interest of few mm^2.

The semi-automated calculation of the nuclei dimensions is an imaging technique that works alongside other investigative techniques in biology such as cytometry and gene expression. The analysis of the images is here carried out with ImageJ for calculating the nuclei sections size and the coordinates x and y for control and mycorrhizal roots. Then a sample of images both mycorrhizal and control are analyzed to compare the sections sizes, the density of the nuclei, the nearest neighbor CE index, the size autocorrelation with variogram and the kriging estimation of the probability of exceeding an arbitrary set threshold.

We can consider the distribution of the size of the nuclei as a marked point process [4] analyzing it with nearest neighbor mathematics, variogram and kriging techniques.

Arbuscular mycorrhizal (AM) symbiosis, a beneficial interaction between the majority of plants and a small group of soil fungi, culminates with the development of arbuscules, structures devoted to the nutrient exchange, inside living root cells. Recent evidence from our research group shed light on the activation of plant cell cycle-related mechanisms during AM colonization using confocal microscopy observations, gene expression and flow cytometry analyses, highlighting the occurrence of endoreduplication in AM colonized areas of the root, with the appearance of polyploid nuclei (greater in size) in arbusculated and neighboring cells of the root cortex [1]. Importantly, due to the correlation between DNA content and nuclear size, the mapping of nuclear ploidy was possible in images of sectioned roots [1].

Considering the previously described results the aim of the present research is to describe the spatial variability of the observed increase in nuclear size in mycorrhizal compared to non-mycorrhizal roots.

The analysis is based on spatial point process and variogram analysis, where the attribute of the point is the equatorial section area of each nucleus expressed in μm^2, and was applied to a large set of 3D confocal z-stack images to define nuclear size variability with ImageJ and the R package gstat [7].

2 Materials and Methods

2.1 Sample Preparation and Confocal Microscopy

All experiments were done using *Medicago truncatula* roots cultivated in Petri dishes. Roots were colonized in-vitro by the fungus *Gigaspora margarita.* Subsequently, 1 cm-long root segments from uninoculated and early-colonized roots were collected. Eight independent samples were collected from independent uncolonized or colonized roots and prepared for confocal imaging.

Root segments were sectioned using a Vibratome and nuclear staining was performed adding the DAPI DNA-specific marker.

Imaging was then performed with a Leica TCS SP2 vertical confocal microscope equipped with a 40X long-range water immersion objective (HCX Apo 0.80), using the 405 nm diode for DAPI excitation. All root sections were scanned at 400 Hz and averaged 2X lines, generating 375 × 375 × 45 μm z stack (z step = 1.5 μm) to be used for image analysis. In Fig. 1, a single frame of a series is displayed.

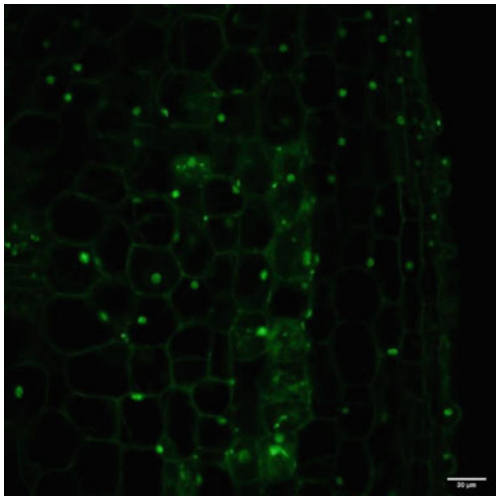


Fig. 1 Frame of a Z-series in confocal imaging of a root. Confocal settings: 375 × 375 × 45 μm (z step = 1.5 μm)

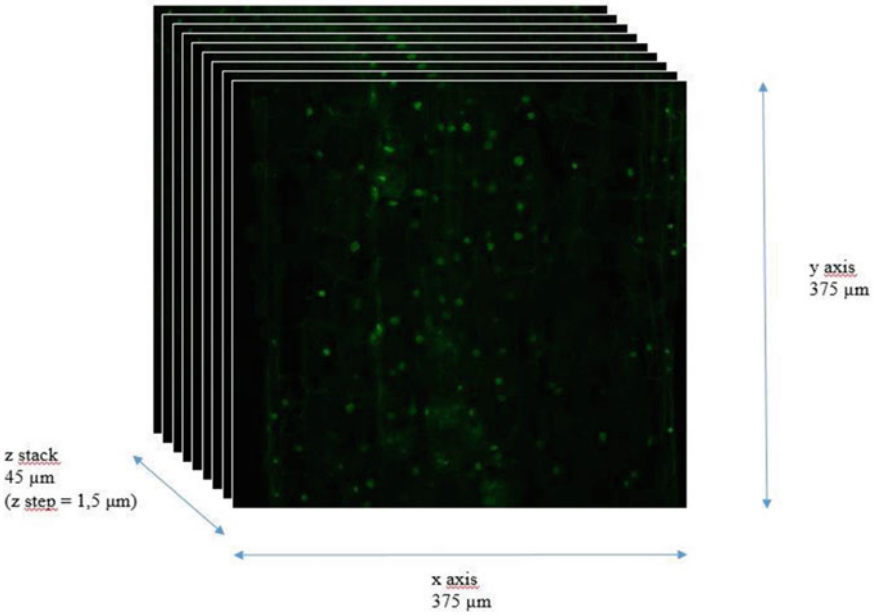


Fig. 2 Z series of confocal imaging of a root. Confocal settings: 375 × 375 × 45 µm (z step = 1.5 µm)

Quantitative microscopy couples digital image analysis programs to an optical technology microscope with laser scan light that describe spots (nuclei sections) as described in Fig. 2, for the measurement of morphology characteristics. The measured nuclei were 3D-detected thanks to a series of frames scanned on the Z axis as described in Fig. 2.

2.2 Image Processing Protocol

The image processing protocol proposed for confocal microscopy analysis in endoreduplication studies includes the following tools:

a. a data set of images from z series divided into 2D analysis frames obtained with the z stacks of the confocal microscope;
b. sample analysis of nuclei dimensions and density,
c. nuclei positioning analysis with Clarks and Evans index;
d. variogram analysis of the nuclear size sections;
e. kriging estimation of probability to exceed the threshold.

The series were collected and the frames were sampled by measuring the size and distance between nuclei in ImageJ with a semi-automatic method. Table 1 describes

Table 1 Nuclei density ρ, mean and standard deviation

Sample	ρ (nuclei/mm^2)	St. dev
Myc	600	73.1
Control	444.6	44.2

the average density value ρ of the nuclei distribution in a frame, and its standard deviation value for the mycorrhizal and control frames.

The positioning index CE [2] expresses how much a distribution of points deviates from a Poisson distribution which is a randomly determined spatial distribution of objects. The average distance between a nucleus section and its nearest neighbor (r_A) is related to the average of the distance expected if the points were randomly distributed, (r_E).

$$CE = \frac{r_A}{r_E} = \frac{\frac{1}{N}\sum_{i=1}^{N} r_i}{0.5\left(\frac{A}{N}\right)^{\frac{1}{2}} + 0.0514\frac{P}{N} + 0.041\frac{P}{N^{\frac{3}{2}}}} \tag{1}$$

r_i is the distance between the nucleus i and its nearest neighbor (in μm), N is the total number of points (nuclei spots) in the image, A is the area of the image in μm^2 and P is the perimeter of the image in μm. For further information about the denominator, see Clarks and Evans [2].

An image with Poisson distribution has a CE value of 1. If we have clustered point clusters, the CE value assumes values smaller than 1. In images with regular distribution points, CE has values greater than 1. To test the index for significant shifts from 1 the statistic proposed by Clarks and Evans [2] is applied. The null hypothesis is tested (H_0: CE = 1 and H_1: CE ≠ 1) using a standard test value normally distributed:

$$c = \frac{r_A - r_E}{\sigma_{r_E}} \tag{2}$$

With

$$\sigma_{r_E} = \frac{0.26136}{\sqrt{N\rho}} \tag{3}$$

where σ_{r_E} is the r_E standard deviation in a poissonian distribution with ρ density.

$$\rho = \frac{N}{A} \tag{4}$$

where N is the number of nuclei in the area A.

Using ImageJ it is possible to semi-automatically measure the size of a spot by transforming the image into an 8-bit image and then automatically contour the nucleus section and calculate the coordinates and the size of the area, as shown in Fig. 3.

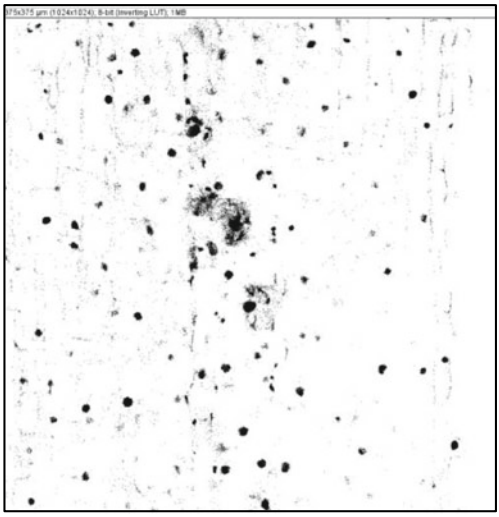


Fig. 3 8 bit image in ImageJ for semiautomated nuclei measurement of coordinates (x, y) and section size (μm^2)

After calculating the positions and dimensions of the nuclei an analysis of the variogram was performed, exploring if there is an autocorrelation for the sections of the nuclei in space.

Semivariance is a function of distance between pairs of objects [3] and the plot of the experimental function is called variogram:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(p_{(i)}, p_{(i+h)}) \epsilon S} \left[y(p_i) - y\left(p_{(i+h)}\right) \right]^2 \tag{5}$$

If we consider digital images from confocal microscopy with resolution $\delta = 2,7307 \frac{pixel}{\mu m}$ as in Fig. 3 the elements of the formula [5] are:

$\gamma(h)$, the semivariance in function of the distance h for a pairs of nuclei sections
$N(h)$, the root nuclei pairs at distance h
S, the frame sample
$y(p_i)$, the nucleus area in μm^2 at the coordinates p_i
$y\left(p_{(i+h)}\right)$, the nucleus area in μm^2 at the coordinates $p_{(i+h)}$.

The sampling design of a series of control and inoculated roots consists of two phases: selection of a root and, subsequently, selection of nuclei in a frame. The point cloud of the variogram and the theoretical estimation function of the variogram are then drawn according to one of the estimation models.

Table 2 Nuclei size, mean and standard deviation

Sample	Nuclei size (μm^2)	St. dev
Myc	37.6	9.8
Control	28.6	11.3

3 Results

3.1 Density and Nuclei Size

The density analysis described a higher spatial frequency of the frames with endoreduplicated nuclei compared to control nuclei; this is also associated with a greater average size of the nuclei as indicated in Tables 1 and 2.

Samples of mycorrhizal nuclei demonstrate a significant difference in nuclear density compared to the control (t-test, p-value < 0.05).

Differences in nuclear size are significant (t test p-value < 0.01).

3.2 CE Index

The CE positioning index showed similar characteristics in both experimental conditions: root nuclei were randomly distributed or showed a tendency to regularity; none of the nuclei distributions were clustered. The fact that mycorrhized images did not have a regular distribution (regular distribution versus random distribution) allowed us to ask whether there is a randomizing effect of something for root nuclei; Is it possible to consider a greater tendency towards the regularity of the distributions of the endoreduplicated nuclei compared to the control ones? We can conclude that a slight shift towards greater regularity in both experimental conditions has occurred, as described in Table 3.

Table 3 Index CE

Series	Nuclei per frame	$\sum \mathbf{d_i}$ (mm) [st.dev]	CE	zTest	Probability
Control 1	55	1.87 [0.48]	1.04	0.61	Not sign
2	63	1.70 [0.25]	1.19	3.09	<0.01
3	61	1.88 [0.06]	1.19	3.04	<0.01
4	70	1.79 [0.60]	1.34	5.77	<0.01
Mycorrhizal 1	75	1.88 [0.14]	1.04	0.75	Not sign
2	89	4.03 [2.09]	1.37	7.13	<0.01
3	96	3.24 [1.11]	2.10	21.6	<0.01
4	76	2.66 [2.10]	1.09	1.73	<0.05

3.3 Estimated Variograms

Spatial plots of the sampled nuclei and estimated variogram are described in Figs. 4 and 5. The value of the sill of the theoretical graph was set on the basis of the variance of the sample measurements while the range was visually evaluated and set equal for the two data series. The nugget was visually evaluated (Table 4).

The experimental variograms are described by the point cloud and the estimated variograms are described by a solid line that draws a theoretical function. The theoretical functions considered are the exponential and the Gaussian. The parameters of the curves in the images of control nuclei are range 30 μm and a semivariance of 35

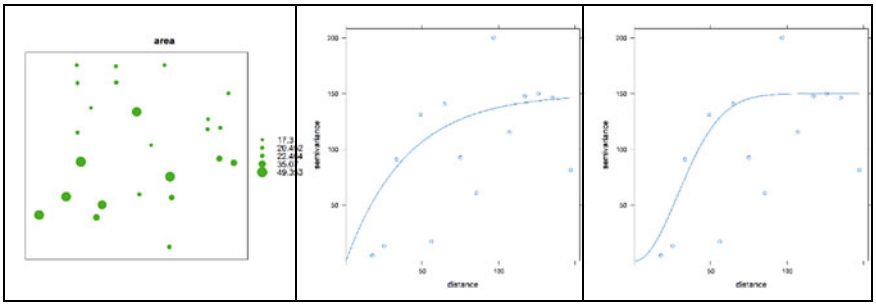


Fig. 4 Sampling area plot of control nuclei and estimated exponential (centre) and Gaussian (right) variograms

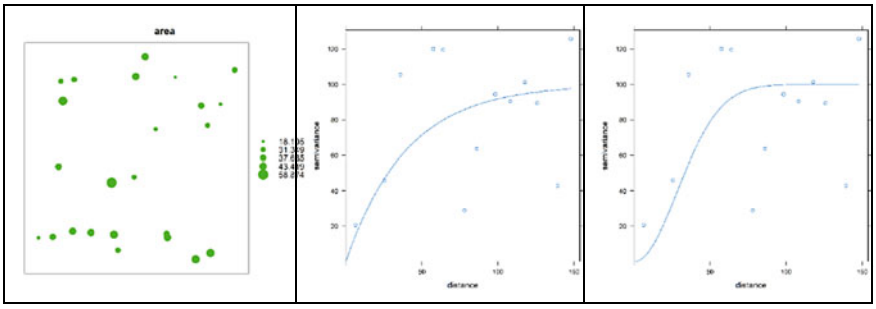


Fig. 5 Sampling area plot of mycorrhizal nuclei and estimated exponential (centre) and Gaussian (right) variograms

Table 4 Parameters of the estimated variogram

	Nugget	Range	Sill
Myc	2	40	50
Control	4	30	35

as sill. The superimposition of estimated and experimental variograms are described in Fig. 4.

Calculation of the experimental variogram and superposition of the theoretical exponential and Gaussian curves for the evaluation of the adaptation in the images of mycorrhizal nuclei as in Fig. 5 describes a range of 40 μm and a semivariance of 50 as sill.

To compare the experimental and the estimated semivariance in function of the distance, the computation of the error of estimates is

$$e_i = Y_i - M_i \tag{6}$$

where

e_i is the error of the estimates for the nucleus i

Y_i is the experimental semivariance

M_i is the model (exponential or Gaussian) estimated semivariance.

Considering the errors for each point i the mean absolute deviation (MAD) is:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{7}$$

And the root mean square error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{8}$$

The Table 5 describes the indexes for control and mycorrhizal digital images groups.

This exploratory research with digital images from confocal microscopy took into consideration spatial statistical analysis techniques used in geostatistics that allow the observation of the variability of study characteristics as a function of spatial coordinates. We drew the point cloud of the variograms for a sample of uninoculated and mycorrhizal roots observing any differences. From this first study using theoretical semivariance estimation models we observed that the estimation models perform better for the frame sample of mycorrhizal roots. This may suggest

Table 5 Performance indexes for the estimation models

Model	MAD	RMSE
Control exponential	3.78	4.73
Control Gaussian	3.61	4.94
Mycorrhizal exponential	2.45	3.20
Mycorrhizal Gaussian	2.70	3.36

minor variability in the distribution of nuclei sizes, to be investigated by further research.

3.4 Estimation with Kriging

Kriging is a modeling technique [3, 6] that uses a linear function to estimate spatial characteristics. We use kriging to estimate the probability of exceeding a diameter threshold value that we arbitrarily set based on the average of the values of the sections in the control roots.

The section of the volume used as threshold is 28 μm^2 mean value of the sections of the nuclei measured in the control frames.

We predict and map the threshold exceedance probabilities in point predictions of thresholding excedaance probabilities over region of interests in the images.

Spatial patterns for the threshold exceedance probabilities are built with confidence regions, estimating exceedance probabilities and standard errors.

Let $D \subset \mathbb{R}^2$ at each location $s \in D$ we assume a spatial point process X_s where and we define threshold exceedance probability

$$\mathbf{P}_{x_0}(s) = P(X_s \geq x_0) \tag{9}$$

where x_0 is 28 μm^2 for any fixed threshold $x_0 \in \mathbb{R}$. Note that $\mathbf{P}_{x_0}(s)$ takes values in [0,1]. We want to estimate $\mathbf{P}_{x_0}(s^*)$ in the notation position $s^* \in D$ where there are supposed to be no observations of nuclei using the estimator $\widehat{P}_s(x_0) = \frac{1}{S}\sum_{i=1}^{S} 1_{\{X_s \geq x_0\}}$ this formula adaptable to arbitrary diagonal points at a distance from the origin of the frame (0,0) and we calculate the probabilities of having nuclei greater than 28 μm^2.

We write the kriging formula for the probability:

$$\widehat{\boldsymbol{P}}_{x_0}\left(s^*\right) = P(X_s \geq x_0) \tag{10}$$

$$\widehat{\boldsymbol{P}}_{x_0}\left(s^*\right) = \Phi\left(\widehat{Q}_{x_0}(s^*)\right) \tag{11}$$

where $\Phi\left(\widehat{Q}_{x_0}(s^*)\right)$ is the normal standard cumulative function

And we calculate with the kriging:

$$\left(\widehat{Q}_{x_0}(s^*)\right) = \beta Y\left(s^*\right) + w\left(s^*\right) \tag{12}$$

where $\beta Y(s^*)$ depends on the grid position of the point and $w(s^*)$ is a function of the second order stationary process described by the theoretical spherical model that we have visually adapted to the experimental variogram and which is described by the spherical model as a function of the distance h between points of the grid:

Table 6 Kriging for 5 notations points

	Distance (±5 μm)	$P(X\_s \geq 28\mu m^2)$ (σ)
Control	75 μm	0.547(0.032)
	150 μm	0.677(0.054)
	240 μm	0.588(0.057)
	330 μm	0.720(0.083)
	405 μm	0.620(0.054)
Mycorrhizal	75 μm	0.471(0.045)
	150 μm	0.683(0.067)
	240 μm	0.718(0.087)
	330 μm	0.640(0.063)
	405 μm	0.532(0.058)

$$\gamma(h) = c_h + \sigma^2\left(\frac{3}{2}\frac{h}{r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right) \tag{13}$$

then we calculate the spherical model as in Eq. (13) where σ^2 is the sill, c_h is the nugget and r is the range.

These parameters of the spherical variogram are estimated by weighted least squares for all grid point in the n frames both of control and mycorrhizal roots.

The kriging predictor is:

$$\left(\widehat{Q}_{x_0}(s^*)\right) = \widehat{\beta}Y\left(s^*\right) + \widehat{w}\left(s^*\right) \tag{14}$$

the kriging as a function of 5 characteristic points described in Table 6 and Figs. 6 and 7 was applied both for the control and for the mycorrhizal obtaining probability areas for dimensions greater than 28 μm^2.

4 Discussion

Our work analyzes the AM root symbiosis, a beneficial interaction between the majority of terrestrial plants and AM fungi. Root cell nuclei participating in this symbiosis can face endoreduplication, a process in which nuclear DNA doubles different times, increasing nuclear size. Our experiment therefore starts from M. truncatula root inoculation with the AM fungus G. margarita and subsequent marking nucleic acid with a fluorophore (DAPI stain), highlighting nuclei to allow observation under a confocal microscope. Thanks to the physical characteristics of the fluorophore and its ability to absorb end emit different light wavelength, we can observe the morphology of cellular objects and proceed with image processing tools. Quantitative microscopy has greatly developed in recent years, allowing the construction

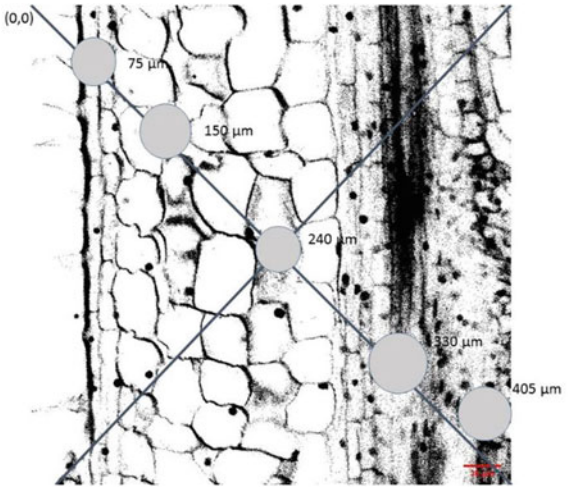


Fig. 6 Control roots. Kriging for 5 spatial points, distances in µm from conventional origin point (0,0). Threshold exceedance probability $P(X\_s \geq 28\mu m^2)$

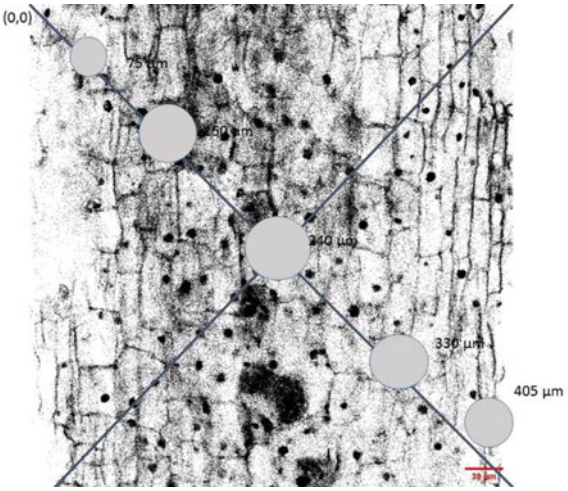


Fig. 7 Mycorrhizal roots. Kriging for 5 spatial points, distances in µm from conventional origin point (0,0); Threshold exceedance probability $P(X\_s \geq 28\mu m^2)$

of measurement methods and comparative statistical tests. The use of a free analysis platform such as ImageJ permits the in-house development of measurement algorithms [1]. The density values were transported in nuclei/mm^2 for greater readability of the results. With a semi-automated analysis procedure of the dimensions, the frames were transformed into 8-bit images and through an image thresholding on pixel intensity (0–255 min, max threshold) it is possible to calculate the nuclear section area. From the results of our sampling on the frames on the Z axis (see Fig. 2) nuclei of mycorrhizal roots seem to have, on average, a larger size compared to those from uninoculated control roots (t-test p-value < 0.01). Also nuclei density in each frame ρ is higher in the mycorrhizal frames, compared to controls (t-test

p-value < 0.05). From our observation we can therefore observe larger nuclei with a higher density in correspondence of mycorrhizal roots, according to endoreduplication events [1]. Our image analysis protocol uses the nearest neighbor mathematics derived from geostatistics, interpreted with the Clarks and Evans index. The analyzes consider the nucleus centers as realizations of a distribution process in the space of the points that can be random, regular or grouped. Results of such analysis are shown in Table 4.

Three quarters of the samples in both controls and mycorrhizal frames showed a CE index over the value of 1, suggesting a nuclear center spatial positioning within sections that do not have a Poissonian random distribution. The comparison of inoculated and uninoculated roots does not allow a useful comparison as both experimental conditions have a predominance of regular distribution in space, according to the CE index. Once these results were obtained, we decided to perform the variogram analysis of nuclear size within frames. In this case, such analysis provides for the calculation of both the nuclear spatial distribution and the correlation of nuclear dimensions at predefined lag distances. The building of the experimental variogram, superimposed with the theoretical exponential and Gaussian variograms, allowed an evaluation of the notable characteristics of the variogram: sill, nugget and range that were initially set using a visual method and inserted in the R gstat environment platform. The results of the geostatistical analyzes transported to confocal microscopy permitted to make comparative evaluations about the inoculated and uninoculated roots. The performance indicators (see Table 5) make it possible to consider the notable models, both exponential and Gaussian, as better superimposable on the experimental curve of the variogram, concerning frames of mycorrhizal roots. This leads us to believe that could be a greater regularity in nuclear spatial distribution, related to nuclear size within mycorrhizal roots.

The estimates obtained with the kriging model also describe the probability to exceed a threshold size in different areas (see Figs. 6 and 7) of the diagonal that virtually divides the root image. Further studies will consider whether the nuclear size distribution is correlated to the proximity of different root tissues (eg. vascular vessels, epidermal or cortical layers).

References

1. Carotenuto, G., Sciascia, I., Oddi, L., Volpe, V., Genre, A.: Size matters: three methods for estimating nuclear size in mycorrhizal roots of Medicago truncatula by image analysis. BMC Plant Biol. **19**(1), 180–193 (2019)
2. Clark, P.J., Evans, F.C.: Distance to nearest neighbour as a measure of spatial relationships in populations. Ecology **35**, 445–453 (1954)
3. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
4. Isaaks, E.H., Srivastava, R.M.: An Introduction to Applied Geostatistics. Oxford University Press, New York (1989)
5. Kint, V., Van Meirvenne, M., Nachtergale, L., Geudens, G., Lust, N.: Spatial methods for quantifying forest stand structure development: a comparison between nearest-neighbor indices and variogram analysis. For. Sci. **49**(1) (2003)

6. Matheron, G.: The Theory of Regionalized Variables and Its Applications. Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu, Paris (1971)
7. Pebesma, E.: Multivariable geostatistics in S: the gstat package. Comput. Geosci. **30**, 683–691 (2004)

Spatially Balanced Indirect Sampling to Estimate the Coverage of the Agricultural Census

Federica Piersimoni, Francesco Pantalone, and Roberto Benedetti

Abstract Coverage error in censuses has become an important statistical issue. In this paper we address the design of the coverage survey of the agricultural census through the use of spatially balanced sampling designs and the employment of indirect sampling framework. Spatially balanced sampling exploits the spatial component of the target population, while indirect sampling is taken into account since a frame linked to the target population is assumed to be used. Following the case of the coverage survey of the agricultural census performed by ISTAT in Italy in 2010, some proposals are presented and their efficiency investigated by means of Monte Carlo simulations. Finally, variance estimation is studied.

Keywords Two-stage design · Local pivotal method · Variance estimation

1 Introduction

Coverage error in censuses is due to omissions or duplications of statistical units in the census enumeration. This has become an important statistical issue. For example, the U.S. Bureau of the Census has been sued in federal court more than 50 times regarding the completeness of the 1980 census. In order to obtain an estimate of the coverage error additional information is necessarily needed. Indeed, an estimate of the coverage error cannot be obtained from the census data themselves, and usually an independent sample survey is obtained over the same target population. This independent survey is often called *coverage survey* (CS) and can either precede (pre-

F. Piersimoni (✉)
Directorate for Methodology and Statistical Process Design, Istat, Rome, Italy
e-mail: piersimo@istat.it

F. Pantalone
Department of Social Statistics and Demography, University of Southampton, Southampton, UK
e-mail: Francesco.Pantalone@soton.ac.uk

R. Benedetti
Department of Economic Studies, "G. d'Annunzio" University, Pescara, Italy
e-mail: benedett@unich.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_27

enumeration survey) or follow (post-enumeration survey) the census. Once the data from the coverage survey are obtained, a model is employed in order to estimate the under/over-count. In agricultural census, the CS is usually employed for the selection of areas, and coverage rate of the agricultural holdings is of interest, i.e. the ratio between the number of agricultural holdings pointed out over the census and the number of really existing agricultural holdings.

In 2010, ISTAT performed the agricultural census with the aim of enumerating all the agricultural holdings in the country. Afterwards, ISTAT carried out the CS aimed at providing a measurement of the degree of coverage of the census with respect to the population of agricultural holdings through an areal sample where the final sampling units were about 1,500 land parcels extracted from the Land Registry Office [14]. It was designed with a two-stage sampling. In the first stage, municipalities were stratified according to their provinces. For each stratum, a number of municipalities were selected with probability proportional to the number of agricultural holdings of the stratum. In the second-stage, land parcels (stored in the *cadastre*) were selected with equal inclusion probabilities from each municipality selected in the first stage. Note that the sampling units are the land parcels of the Land Registry.

In this work, following the set-up of ISTAT, we investigate the designing of the CS for agricultural census through spatially balanced sampling in the context of indirect sampling. The use of the former is motivated by the strong spatial component of the application. Indeed, units distributed over a region of interest tend to be similar since they are influenced by the same set of factors, which is especially true in agricultural surveys, where units close together are influenced by the same soil fertility, weather, pollution, and other spatial factors. In order to exploit this feature, spatially balanced sampling designs select sample well spread over the region of interest. For a review, see [3]. The framework of indirect sampling [16] is taken into consideration, since in this set-up we sample land parcels in order to select agricultural holdings. Indeed, we do not sample from a frame of the target population (of agricultural holdings), but we sample from a frame (of land parcels) linked to the target population. The paper is organized as follows. In Sect. 2 we introduce the design-based approach and the theoretical framework of indirect sampling, while in Sect. 3 we briefly review the current literature of spatially balanced sampling designs. In Sect. 4 we introduce and discuss some proposals for the CS, and we evaluate their efficiency by means of a Monte Carlo simulation. Finally, Sect. 5 is dedicated to variance estimation and Sect. 6 provides conclusions and future research.

2 Theoretical Framework

In this section we introduce the design-based paradigm and the indirect sampling framework. The notation strictly follows the one in [16]. Given a finite target population $U^B = \{1, \ldots, M^B\}$, we suppose the target of interest is the total of the variable of interest y, that is

$$t_y = \sum_{i=1}^{M^B} y_i. \tag{1}$$

In the design-based framework the variable of interest y is considered fixed and the only source of randomness comes from the selection of a sample s from the population U^B by means of a sampling design $p(s)$. We denote with $\mathcal{S}$ the set of all possible subsets of U^B and we note that a sample without replacement is an element $s \in \mathcal{S}$. A sampling design is a probability distribution on $\mathcal{S}$ such that $p(s) \geq 0$ and $\sum_{s\in\mathcal{S}} p(s) = 1$. The sampling design defines in turn the probability of selecting unit i, which is called first-order inclusion probability and given by $\pi_i = \sum_{s\ni i} p(s)$, and the probability of selecting unit i and j in the same sample, which is called second-order inclusion probability and given by $\pi_{ij} = \sum_{s\supset\{i,j\}} p(s)$. We can estimate the total t_y through the [12] estimator (HT)

$$t_{y,HT} = \sum_{i\in s} \frac{y_i}{\pi_i}, \tag{2}$$

which is unbiased with respect to the design when $\pi_i > 0 \; \forall \, i \in U^B$.

Standard way to proceed in survey sampling requires a frame for the target population U^B, from which a sample of units is selected by means of a sampling design $p(s)$. Unfortunately, in some circumstances no frame is available for U^B. We may have a frame for another population $U^A = \{1, \ldots, M^A\}$, which is linked to the target population U^B. Indirect sampling consists in selecting a sample s^A from U^A and exploit the linkage with U^B in order to produce the desired estimate. The major challenge is to assign a selection probability or an estimation weight to the units of the population U^B. In this context, a well-known method is the generalized weight share method [15], or GSWM, which is a generalization of the weight share method [7] and produces an estimation weight for each surveyed unit of the population U^B through an average of the sampling weights of the population U^A. As starting point, the GSWM supposes that

(i) the population target U^B is divided into N clusters, where the ith cluster has M_i^B units;
(ii) there exists a relationship between unit j of population U^A and unit k of cluster i of the population U^B. This relationship is indicated through an indicator variable $l_{j,ik}$, which is equal to 1 when a link exists between $j \in U^A$ and unit $ik \in U^B$, and equal to 0 otherwise.

Let us define the following quantities

$$L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} \qquad \text{and} \qquad L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik},$$

where L_j^A gives the number of links between the unit $j \in U^A$ and the units k of cluster i of population U^B, and L_{ik}^B gives the number of links for any unit k of a cluster i of population U^B. Note that in order to use the GSWM, we must satisfy $L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} > 0$. We select a sample s^A of m_A units from U^A by means of a sampling design with first order inclusion probabilities $\pi_j^A > 0 \, \forall j \in U^A$. Once the sample s^A is selected, for each unit $j \in s^A$ we identify the units ik of U^B that have a non-zero link with j, and for each of those units we assume we can set up the list of M_i^B units of cluster i containing this unit. Therefore, each cluster i represents itself a population U_i^B where $U^B = \bigcup_{i=1}^{N} U_i^B$. Then, let Ω^B be the set of n clusters identified by the units $j \in s^A$, that is $\Omega^B = \{i \in U^B | \exists \; j \in s^A \text{ and } L_{i,j} > 0\}$ with $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$.

We can now survey all the units k of cluster $i \in \Omega^B$. In particular, we record the variable of interest y_{ik} and the number of links L_{ik}^B. We can rewrite the target of inference (1) as

$$Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik},$$

which can be estimated by

$$\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik}, \tag{3}$$

where n is the number of clusters surveyed and w_{ik} is the weight assigned to unit k of cluster i. These weights are obtained by the GWSM, in such a way the final weights are calculated according to a weighted method within each surveyed cluster. Indeed, for the weight w_{ik}, the method starts with the computation of an initial weight, which is the inverse of the inclusion probability of the unit selected from U^A and with link with the cluster i (if the unit does not have a link, the initial weight is equal to zero). Then, the final weight is obtained as the ratio of the sum of the initial weights for the cluster over the total number of links in the cluster, and it is assigned to all units in the cluster.

3 Spatially Balanced Sampling Design

Units that are spatially distributed usually show spatial dependence. With regard to agricultural research, surveys are routinely used to gather data. The observed units are typically geo-referenced and therefore sampling designs that take into account such information could be employed. Indeed, from the geographical position of each unit i we can derive, according to some distance definition, a matrix that specifies how far all the pairs of units in the population are. This is easily applicable to points, since simple concepts of distance between set of coordinates can be applied, and needs

some adaption when other elements are used, for example if we need to sample polygons. In this case we either identify polygons with their centroids, or we should use as a distance the notion of contiguity between areal units.

When it comes to select units from a spatial population, several methods have been introduced in the sampling literature. Among different proposals, a new group of sampling designs has emerged lately and referred to as *spatially balanced sampling designs*. This type of design selects samples well spread over the population of interest. The idea is that a spread sample could capture the spatial heterogeneity of the population, which in turn could improve the efficiency of estimates compared to the efficiency of estimates achieved by data obtained from non-spatial sampling design (for more details about efficiency and extensive simulations, see [4]). This idea is supported by theoretical reasons. In particular, [11] postulated a model for the spatial dependence of the U^B and exploited the concept of *Anticipated Variance* (AV) [13]. Indeed, suppose the following linear model holds for each unit of U^B

$$\begin{cases} y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \\ E_m(\epsilon_i) = 0 \\ C_m(\epsilon_i, \epsilon_j) = \sigma_i \sigma_j \rho_{ij} \end{cases} \tag{4}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, ϵ_i is a zero-mean random variable with variance σ_i^2, ρ_{ij} is an autocorrelation parameter for $i \neq j$ and such that $\rho_{ij} = 1$ if $i = j$, and E_m and C_m denote expectation and covariance with respect to the model, respectively,

The AV of the HT estimator of the total y under the model (4) is given by

$$\begin{aligned} AV(t_{y,HT}) = E_s E_m (t_{y,HT} - t_y)^2 = E\left[\left(\sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i=1}^{N} \mathbf{x}_i\right)^\top \boldsymbol{\beta}\right]^2 \\ + \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}, \end{aligned} \tag{5}$$

where E_s denote expectation respect to the design.

The second term in (5) refers to the dependence of the observed units. Indeed, if we assume that ρ_{ij} decreases as the distance d_{ij} increases, then selecting units far apart (therefore altering the second-order inclusion probabilities π_{ij}) reduces the AV of the estimates. Therefore, use of a spatially balanced sampling can reduce the uncertainty of our estimates.

One challenge in the development of these sampling designs is due to the fact that, while the second order inclusion probabilities can vary during the selection mechanism (in order to spread the sample the closer two units are, the less likely they need to be selected), the first order inclusion probabilities are fixed in advance and therefore they cannot vary. To date, different spatially balanced sampling designs have been introduced in the literature. To name a few, the Generalized Random

Tessellation Stratified (GRTS) design [17] selects samples with fixed size n, mapping the two-dimensional spatial population into one dimension, while preserving some spatial relationships between the units. The Spatially Correlated Poisson Sampling (SCPS) design [8] is a modification of Correlated Poisson Sampling [5] and is based on a list sequential criterion of random decisions. The Local Pivotal Method (LPM) [9] is a derivation of the Pivotal Method [6], which consists in updating at each step the inclusion probability for two units so that the sampling outcome is decided for at least one of the two units, and the Product Within Distance [2], which is an MCMC-based method that uses a summary index of the distance matrix. For a detailed review, see [3] and [4].

Among these sampling designs, we choose to use the LPM in the simulations of next section. Therefore, we explain here how it works. This design is a derivation of the Pivotal Method [6], which at each step chooses two random units and updates the inclusion probabilities for those units so that the sampling outcome is decided for at least one of them (i.e. one of them is surely included in the sample). Once a unit is selected, it cannot be chosen again in the following steps, and the procedure continues until the sampling outcome is decided for each unit in the population. Therefore, the sample is obtained in at most N steps. In particular, let i and j be the units chosen at stage t, with corresponding inclusion probabilities π_i and π_j. If $\pi_i + \pi_j < 1$, the inclusion probabilities are updated as follows

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

otherwise, if $\pi_i + \pi_j \geq 1$,

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{1-\pi_j}{2-\pi_i+\pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{1-\pi_i}{2-\pi_i+\pi_j} \end{cases}$$

In order to achieve a spread sample, the LPM adapts the Pivotal Method procedure. Indeed, instead of randomly selecting units at each step, the LPM selects nearby units. This sampling design allows to select well-spread samples with equal or unequal inclusion probabilities, and with fixed sample size n as long as the first-order inclusion probabilities sum up to n. Moreover, there are two versions of this method, which differ on the way the nearby units are selected at each step and are referred to as LPM 1 and LPM 2. The first one achieves samples that are more spatially balanced than the samples obtained by the second one, but with a larger expected number of computations (for LPM 1 it is, in the worst case, proportional to N^3, and at best proportional to N^2, while for LPM 2 it is proportional to N^2).

4 Proposals and Simulations

In this section we discuss some proposals for the sampling design to employ in the CS, and we present a set of simulations where the efficiency of these proposals is investigated. Indeed, following the notation introduced in the previous sections, we investigate the use of spatially balanced sampling designs for the selection of units from the population U^A in order to estimate the total t_y of the population of interest U^B. This is done by employing the GSWM. We do this following the set-up of the real-case scenario of the Italian agricultural census carried out by ISTAT, and the following CS employed with the aim of estimating the degree of coverage that the census provided with respect to the population of agricultural holdings [14]. In particular, the CS was performed by means of a two-stage sampling, where in the first stage municipalities were firstly stratified according to their provinces and then selected with probability proportional to the number of agricultural holdings of the stratum. In the second stage, land parcels (stored in the *cadastre*) were selected with equal inclusion probabilities from each municipality selected in the first stage. Finally, the agricultural holdings were selected from the land parcels. See Fig. 1 for a picture of the hierarchical levels. From the above description, we can see that ISTAT actually performed an indirect sampling: land parcels were the sampled units, which were linked to the agricultural holdings. Therefore, in terms of our notation, U^A represents the linked population of land parcels while U^B represents the actual target population of agricultural holdings.

In this work, we wanted to simulate this situation and investigate the role that sampling designs, and particularly spatially sampling designs, play in the efficiency of the estimates of the census coverage. For this reason, we firstly simulated an artificial population that mimics the aforementioned situation, and then we employed a Monte Carlo simulation where the impact of the sampling designs are evaluated by means of Root Mean Squared Error (RMSE).

The artificial population is a grid made up by land parcels laid over municipalities (see Fig. 2) and is generated through the following steps.

1. A regular 30×30 grid that represents the land parcels is generated, and the municipality limits are generated by aggregation of the land parcels by means of a clustering algorithm based on a Minimum Spanning Tree [1], which is used in 50 contiguous groups (clusters) and with a number of prefixed land parcels between 12 and 38.
2. Two populations that represent clustered and sparse configurations of agricultural holdings, respectively, are generated. Toward this end, two series of 1,000 points are generated according to the Neyman-Scott process with Cauchy cluster kernel [18] and are overlaid on the grid. These points represent the center position of the agricultural holdings. The intensity of the Poisson process is equal to 10, and the mean number of units per cluster is 100. Two different scale parameters 0.05 and 0.1 are used for the two series, respectively, where the former is used for the clustered population while the latter for the sparse population.

Fig. 1 Hierarchical levels for the Italian agricultural census. Thick lines delimit regions, thin lines delimit provinces

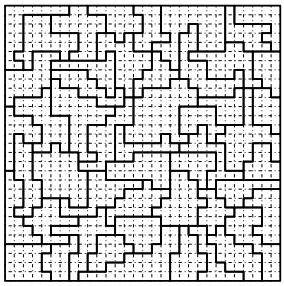


Fig. 2 Artificial population

3. A size is assigned to each agricultural holding according to a negative exponential distribution of parameter $\beta = [0, 2]$. The sizes are approximated to the upper integer.
4. A variable [0, 1] (censused/not censused) is generated through a Markov Chain Monte Carlo (MCMC) algorithm such that the probability of being 1 is simultaneously inversely proportional to the size and proportional to the frequencies of 1 in the neighborhood (spatial dependence). The frequencies are fixed to 850 and 150, for the 0 and 1, respectively.
5. Three sets of links are generated with the population of the land parcels (i.e. in which land parcels the company has land in use). After removing the land parcels where the business center is located, links are generated (size-1) with probability inversely proportional to the distance between the agricultural holding and the land parcels, raised to a control parameter set equal to 2, 3, 10 to ensure that these links are more or less probable increasing the distance.

In this analysis, we are interested in understanding if (i) spatially balanced sampling design provides gain in efficiency of the final estimates in the context of indirect sampling, and (ii) how a one-stage design compares to a two-stage design. Indeed, we do know that spatially balanced sampling designs could provide gain in efficiency in standard spatial population set-up (as we outlined in Sect. 3 and references therein), but the indirect sampling framework used here introduces the element of the linkage. The choice between the one-stage and two-stage design is not trivial: we believe that the former can provide the greatest gain in efficiency, but we do recognize that the latter is employed for practical, administrative, and economical reasons.

Toward this end, a Monte Carlo simulation is employed, with $M = 10{,}000$ replicated samples of land parcels of size $n = \{9, 24, 45\}$ selected from the aforementioned populations, and in different scenarios characterized by the different sets of links (L^2, L^3, and L^{10}) generated as explained in the previous step 5. In the two-stage sampling designs, the first-stage is employed to select municipalities, and the second-stage selects land parcels. Therefore, for $n = 9$ three municipalities and three land parcels per municipality are selected, for $n = 24$ six municipalities and four land parcels per municipality are selected, and for $n = 45$ nine municipalities and five land parcels per municipality are selected. From the land parcels, every agricultural holding linked to them is then selected. The estimator (3) is employed, and the proposals are compared in terms of the Monte Carlo RMSE. As we anticipated in the previous section, in this set of simulations we choose to use the LPM for the spatially balanced sampling design to employ. This is because the efficiency of this sampling design is close to the SCPS and PWD [4]. In particular, the sampling designs employed and evaluated are:

I. one-stage design by means of SRS (as benchmark), referred to as SRS.
II. One-stage design by means of LPM, referred to as LPM.
III. Two-stage design by means of SRS (as benchmark), referred to as 2S SRS.
IV. Two-stage design by means of LPM, referred to as 2S LPM.

Table 1 summarizes the results. Table 1a shows that the use of spatially balanced sampling design provides an improvement in terms of RMSE, both with a one-stage and two-stage designs. Moreover, efficiency increases when the population is clustered and the links are stronger. Table 1b indicates that, when possible, a one-stage design should be preferred to a two-stage design. Since this is not always possible, when the two-stage design needs to be employed, the spatial version should be adopted.

5 Variance Estimation

This section presents a preliminary study on variance estimation. We start off the analysis with the duality theorem of [16]. Let $z_{ik} = Y_i / L_i^B$ where $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$, and $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B \; \forall k \in U_i^B$. The estimator $\hat{Y}^B$ can also be written as

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \tag{6}$$

where $Z_j = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{i,jk} z_{ik}$. The duality theorem allows us to construct the new variable Z and perform variance estimation by means of the standard literature of the HT variance estimation.

Table 1 Relative RMSE for the different scenarios

		Sparse population			Clustered population		
	n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
(a) Relative RMSE							
LPM vs SRS	9	0.988	0.990	0.985	0.929	0.919	0.928
	24	0.959	0.939	0.942	0.878	0.858	0.855
	45	0.935	0.903	0.894	0.838	0.817	0.806
2S LPM vs 2S SRS	9	0.954	0.939	0.927	0.909	0.899	0.894
	24	0.950	0.935	0.920	0.880	0.872	0.878
	45	0.938	0.925	0.915	0.883	0.867	0.870
(b) Relative RMSE							
2S SRS vs SRS	9	1.229	1.258	1.256	1.156	1.157	1.181
	24	1.293	1.316	1.330	1.238	1.260	1.273
	45	1.340	1.371	1.385	1.266	1.311	1.317
2S LPM vs LPM	9	1.186	1.192	1.183	1.131	1.133	1.138
	24	1.280	1.310	1.299	1.240	1.280	1.308
	45	1.344	1.406	1.416	1.335	1.392	1.423

Grafström and Schelin [10] propose to estimate the variance of the HT estimator under the LPM design by using:

$$\hat{V}(\hat{Y}^B) = \frac{1}{2} \sum_{j \in s^A} \left(\frac{Z_j}{\pi_j^A} - \frac{Z_{j(i)}}{\pi_{j(i)}^A} \right)^2 \tag{7}$$

where $j(i)$ is the nearest neighbor to i in the sample. For a two-stage sampling design, the following applies

$$\hat{V}(\hat{Y}^B) = \hat{V}_1(\hat{Y}_k^B) + \frac{N_1}{n_1} \hat{V}_2(\hat{Y}_j^B), \tag{8}$$

where k and j are indices that run respectively on the sample municipalities and the portions of map selected in the second stage sample within each municipality, and $\hat{V}_1$ and $\hat{V}_2$ indicate the estimated variance of the first and second stage, respectively. Therefore, we propose to use

$$\hat{V}(\hat{Y}^B) = \left[\frac{1}{2} \sum_{j \in s_1^A} \left(\frac{\hat{Z}_j}{\pi_j^{A_1}} - \frac{\hat{Z}_{j(i)}}{\pi_{j(i)}^{A_1}} \right)^2 \right] + \frac{N_1}{n_1} \sum_{j \in s_1^A} \left[\frac{1}{2} \sum_{k \in s_2^A, k \in j} \left(\frac{Z_k}{\pi_k^{A^2}} - \frac{Z_{k(h)}}{\pi_{k(h)}^{A_2}} \right)^2 \right] \tag{9}$$

Table 2 Bias and RMSE of the variance estimator (7). Case of LPM

		Sparse population			Clustered population		
	n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
Bias	9	−16.94	−20.42	−22.19	−13.93	−17.37	−19.65
	24	−2.93	−2.01	−1.66	−0.07	1.86	3.31
	45	0.61	1.63	2.59	2.45	3.91	5.39
RMSE	9	61.03	71.05	76.60	66.93	82.86	90.86
	24	25.50	29.60	31.79	27.66	33.21	35.39
	45	14.09	16.25	17.30	15.17	17.73	18.82

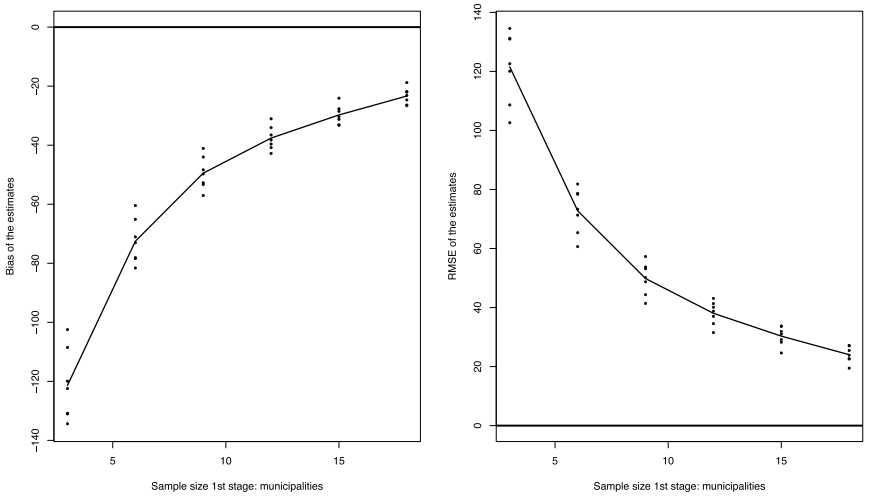


Fig. 3 Bias and RMSE of the variance estimator (9). Case of two-stage design

As a preliminary study, we run a Monte Carlo simulation of $M = 10,000$ replicates for the case of one-stage LPM, where estimator (7) is used. Results are reported in Table 2 and suggest that we should pay attention when the sample size is really small. Indeed, in this case the bias is substantial and the use of this estimator in this particular setting is not recommended. This bias decreases when the sample size increases, as we would expect. We also run a Monte Carlo simulation of $M = 10,000$ replicates for the case of the two-stage LPM. In this case, the estimator (9) is employed. Figure 3 summarizes the results. This variance estimator suffers from bias when (i) the sample sizes of the two stages n_1 and n_2 are very small, and (ii) the sampling fractions of the two stages f_1 and f_2 approach 0.50, since in this case we do not have spreading of the units anymore. These are preliminary results, which suggest when we should avoid the use of such estimator. A deeper study of the variance estimation is a current topic of research.

6 Conclusion

In this paper we focused on the CS of agricultural census. Indeed, we proposed the use of spatially balanced designs while the framework of indirect sampling is employed. Results show that the combination of these two methods allows to achieve good results in terms of RMSE, which is investigated by means of Monte Carlo simulation. In particular, spatial sampling design provides a gain in efficiency in terms of RMSE. A preliminary study of the variance estimation has been presented as well, and some indications have been drawn. A deeper study of this aspect is a current topic of research. Moreover, future research will involve use of real data.

Acknowledgements Federica Piersimoni's work represents her views and does not necessarily reflect those of ISTAT. This work has been presented at the *Advisory Committee on Statistical Methods* of ISTAT on 12 January 2021.

References

1. Assunção, R.M., Neves, M.C., Câmara, G., da Costa Freitas, C.: Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. Int. J. Geograph. Inf. Sci. **20**(7), 797–811 (2006)
2. Benedetti, R., Piersimoni, F.: A spatially balanced design with probability function proportional to the within sample distance. Biometrical J. **59**(5), 1067–1084 (2017)
3. Benedetti, R., Piersimoni, F., Postiglione, P.: Sampling Spatial Units for Agricultural Surveys. Advances in Spatial Science Series. Springer, Berlin, Heidelberg (2015)
4. Benedetti, R., Piersimoni, F., Postiglione, P.: Spatially balanced sampling: a review and a reappraisal. Int. Stat. Rev. **85**(3), 439–454 (2017)
5. Bondesson, L., Thorburn, D.: A list sequential sampling method suitable for real-time sampling. Scand. J. Stat. **35**(3), 466–483 (2008)
6. Deville, J.-C., Tillè, Y.: Selection of several unequal probability samples from the same population. J. Stat. Plann. Infer. **86**(1), 215–227 (2000)
7. Ernst, L.R.: Weighting Issues for Longitudinal Household and Family Estimates. US Bureau of the Census (1986)
8. Grafström, A.: Spatially correlated poisson sampling. J. Stat. Plann. Infer. (2011)
9. Grafström, A.: Spatially balanced sampling through the pivotal method. Biometrics **68**(2), 514–521 (2012)
10. Grafström, A., Schelin, L.: How to select representative samples. Scand. J. Stat. **41**(2), 277–290 (2014)
11. Grafström, A., Tillè, Y.: Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics **24**(2), 120–131 (2013)
12. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**(260), 663–685 (1952)
13. Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. J. Am. Stat. Assoc. **77**(377), 89–96 (1982)
14. ISTAT: Atti del 6° censimento generale dell'agricoltura, la valutazione della qualità, fascicolo 5. Technical report, ISTAT (2013)
15. Lavallée, P.: Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. Surv. Methodol. **21**(1), 25–32 (1995)
16. Lavallée, P.: Indirect Sampling. Springer, New York (2007)

17. Stevens, D.L., Olsen, A.R.: Spatially balanced sampling of natural resources. J. Am. Stat. Assoc. **99**(465), 262–278 (2004)
18. Waagepetersen, R.P.: An estimating function approach to inference for inhomogeneous Neyman-Scott processes. Biometrics **63**(1), 252–258 (2007)

The Assessment of Environmental and Income Inequalities

Michele Costa

Abstract We analyze the income and environmental inequalities and their interrelation by means of both the multidimensional poverty measurement and the Gini index decomposition. We stress how overlap between the two dimensions plays a relevant role and allows powerful insights on the contribution of the environmental dimensions to poverty. Our finding underlines that bad environmental conditions are more likely among low-income units and represent a relevant inequality factor.

Keywords Environmental inequality · Income inequality · Overlap

1 Introduction

The growing attention and the widespread concern for the SDGs underpin the relevance of the interconnession between the 13 goals, which requires the harmonisation and the coordination of the variety of actions implemented at national and international level. In this work we address three different SDGs and their interaction: the 1st, related to overall inequality, the 10th, related to economic inequality, and the 13th, related to environmental inequality.

Environmental inequality refers to unequal distribution of opportunities related to environment: environmental degradation does not affect everyone equally and the effects of environmental risks are not uniformly distributed. Moreover, environmental inequality has a strong impact on the economic and social system for a wide range of reasons, both ethical, normative and economic.

There is a natural correlation between income and environmental inequalities [9], two of the main dimensions of poverty. In this paper we aim to tackle the interplay between these two dimensions [2], which is a key issue in inequality analyses and can be considered as a major threat to economic resilience [1].

We contribute to the analysis of the interplay between income and environmental dimensions by building on both the multidimensional measurement of poverty and

M. Costa (✉)
Department of Economics, University of Bologna, piazza Scaravilli 2, Bologna, Italy
e-mail: michele.costa@unibo.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_28

the decomposition of the Gini index. In particular, in the analysis of the income and environmental inequalities, we underline the role of overlap existing between the two dimensions.

There is a wide range of environmental risks and, depending on the definition which we adopt for the environmental dimension [8], results and policy implications change, while usually confirming that bad environmental conditions are more likely among low-income units, be they individuals, households, communities or even geographical partitions [5]. Furthermore, sustainability policies interact and deal with many different subjects but, in order to be effective, are characterized and united by the need to be data based. We contribute to answer to this need and the methods we are going to present are extremely flexible and can be adapted to different definitions and the analysis of multiple risks.

2 Methodology

The most widespread measure of inequality, the Gini index, has already been successfully used in the study of environmental inequality, both in its traditional expression and in an environmental Paglin-Gini extension proposed by [10] or a spatialized Gini index applied to environmental segregation [11].

A relevant source of information about interplay between income and environmental dimensions is represented by overlap existing between income distributions related to low and high quality of environmental conditions. In the absence of overlap, environmental dimension fully explains income inequality, while, on the contrary, a perfect overlap suggests that environmental and economic dimensions are independent.

In order to evaluate the degree of the overlap we have two measures at our disposal, both introduced by Gini: the probability of transvariation, which measures the frequency of overlapping occurrences, and the intensity of transvariation, which evaluates the extent of the overlap.

Beside probability and intensity of transvariation, we can also take into account the overlap within the Gini index, through its decomposition. In particular, we refer to the decomposition proposed by Dagum [7], which has among its strengths the role attributed to overlap.

A preliminary step to the joint analysis of the economic and environmental dimensions is represented by the traditional unidimensional poverty representation. Given a population of n units, it is developed on a vector of incomes, $\mathbf{y} = (y_1, y_2, ..., y_n)$ and on a poverty line z_y on the basis of which we define a zero-one vector $\mathbf{g_y}$ where

$$g_{yi} = \begin{cases} 1 \text{ if } y_i < z_y \\ 0 \text{ otherwise} \end{cases}$$

The number of poor units with respect to income, q_y, is the sum of the vector $\mathbf{g_y}$.

By adding the environmental dimension, a vector $\mathbf{e} = (e_1, e_2, ..., e_n)$ representative of the environmental conditions is introduced, together with a poverty line z_e on the basis of which we define a zero-one vector $\mathbf{g_e}$ where

$$g_{ei} = \begin{cases} 1 \text{ if } e_i < z_e \\ 0 \text{ otherwise} \end{cases}$$

The number of poor units with respect to the environmental dimension, q_e, is the sum of the vector $\mathbf{g_e}$. The use of a z_e threshold allows for the treatment of many different environmental risks and is robust with respect to various definitions of the environmental dimension.

In the following we refer to the case of only two groups, poor and non-poor, for each of the two dimensions, but it is possible to generalize to the case of more than two groups, allowing different levels or intensities of poverty, a situation that we would like to address in a future work.

The joint analysis of the two dimensions leads to classify the n units of the population in four classes which can be reported in the 2×2 Table 1, where we find the q poor units in both dimensions, the $q_e - q$ units which are poor with respect environmental dimension but not according to income, the $q_y - q$ units poor only with respect to income, and the $n - q_e - q_y - q$ units that are not poor for both dimensions.

In absence of overlapping, we have

$$q_e = q_y = q,$$

which implies both

$$q_e - q = q_y - q = 0 \text{ and } n - q_e - q_y + q = n - q :$$

poor (non poor) units according to income are also poor (non poor) units for the environment.

On the other side, when the overlap is perfect, the conditional distributions are the same, that is

$$q_y/n = q/q_e = (q_y - q)/(n - q_e).$$

Overlap becomes a critical issue since, when it is absent, the groups of the poor identified by the two dimensions are coincident and, therefore, the environmental dimension fully explains economic inequality, and vice versa. As overlap increases, we have a weaker influence of environmental dimension on economic inequality, with the minimum effect corresponding to the case of perfect overlap. Notwithstanding the relevant information provided by the analysis by column of Table 1, it is its analysis by row that results essential for the analysis of overlap and that can be carried out by means of both unidimensional and multidimensional indicators.

Table 1 Poor and non-poor units by income and environmental conditions

		Income		
		Poor	Non poor	
Environment	Poor	q	$q_e - q$	q_e
	Non poor	$q_y - q$	$n - q_e - q_y + q$	$n - q_e$
		q_y	$n - q_y$	n

2.1 Unidimensional Indicators

The simplest indicators that analyze the two dimensions separately are the head count ratios

$$H_y = \sum_{i=1}^{n} g_{yi}/n = q_y/n \text{ and}$$

$$H_e = \sum_{i=1}^{n} g_{ei}/n = q_e/n.$$

In the case (such as the SHIW data analyzed in the next Sect.) of sample surveys, it is necessary to also include the weight a_i attached to each sample unit, thus obtaining

$$H = \sum_{i=1}^{n} g_i a_i / \sum_{i=1}^{n} a_i.$$

The head count ratio can be computed not only for marginal distributions, but also with respect to conditional ones: the head count for income conditionally to environmental poor is

$$H_{y|ep} = q/q_e$$

and, analogously, we have

$$H_{y|enp} = (q_y - q)/(n - q_e),$$

$$H_{e|yp} = q/q_y \text{ and}$$

$$H_{e|ynp} = (q_e - q)/(n - q_y).$$

In case of perfect overlap, given equal conditional distributions, we get

$$H_y = H_{y|ep} = H_{y|enp}$$

while an increasing overlap leads to an increasing difference between head count ratios related to marginal and conditional distributions.

A further widely known and widely used unidimensional indicator is the Gini index G, which, similarly to the head count ratio, can be calculated on the n total observations, but also on the conditional distributions: with reference, for example, only to units poor for income, we obtain G_{py}, conditionally to units poor for the environment we have G_{pe}, etc.

Overall, we get a set of measures which allow to thoroughly evaluate and compare the inequality levels of different groups of units.

A last relevant unidimensional poverty indicator is the Sen index [12], which can be expressed as

$$S = H_y(I_{py} + (1 - I_{py})G_{py})$$

where I_p is the mean over the poor of the normalized poverty gap,

$$I_{py} = \frac{1}{q_y} \sum_{i=1}^{q_y} \left(\frac{z_y - y_i}{z_y} \right)$$

and G_{py} is the Gini index of the poor.

Sens's proposal is based on the three *Is*, i.e., the three key elements of poverty, its size, H_y, its depth. I_{py}, and its distribution among the poor, G_{py}, and allows powerful insight on different aspects of poverty.

2.2 Bidimensional Indicators

From H_y and H_e a two-dimensional indicator can be easily derived as a weighted average of the unidimensional measures:

$$H_{ye} = (H_y w_y + H_e w_e)/(w_y + w_e).$$

Among the possible weighting structures we refer to

$$w_y = log(n/q_y) \text{ and } w_e = log(n/q_e),$$

following the proposal by Cerioli and Zani [4] which aim to measure the intensity of deprivation and social exclusion related to each dimension. The more q_e is less than q_y, the more w_e is greater than w_y (or viceversa), thus indicating the different level of social exclusion attached to the two dimensions investigated in the paper. Within this framework, $w_y = w_e$ if and only if $q_y = q_e$, that is if $H_y = H_e$.

Alternatively, the simplest weight structure implies an equally-weighted multidimensional index:

$$w_y = w_e = 0.5.$$

Beyond the value of the indicator H_{ye}, the relevant aspect in the choice of the weighting system is the different composition of the set of poor units determined by different weights.

In order to analyze the interplay between environmental and income dimension the three indices H_y, H_e and H_{ye} are extremely helpful since, on their basis, it is possible to compare the three sets of poor units identified by them.

The joint analysis of the two dimensions can also be obtained by dividing the population in two subgroups, the first with the q_e poor units and the second with $(n - q_e)$ non poor units according to the environmental conditions, and deriving the Gini index for income as

$$G = G_{pe}p_{pe}s_{pe} + G_{npe}p_{npe}s_{npe} + G_{pe.npe}p_{pe}s_{npe} + G_{npe.pe}p_{npe}s_{pe}$$

where

$p_{pe} = q_e/n$ and $p_{npe} = (n - q_e)/n$ indicate the population shares,

$s_{pe} = p_{pe}\bar{y}_{pe}/\bar{y}$ and $s_{npe} = p_{npe}\bar{y}_{npe}/\bar{y}$ the income shares,

G_{pe} and G_{npe} the Gini indices of the two subgroups, and

$G_{pe.npe}$ is the Gini index between the poor and non-poor units according to the environmental conditions, with $G_{pe.npe} = G_{npe.pe}$ and

$$G_{pe.npe} = \frac{1}{q_e(n - q_e)(\bar{y}_{pe} + \bar{y}_{npe})} \sum_{i=1}^{q_e} \sum_{r=1}^{n-q_e} |y_{pei} - y_{nper}|.$$

The Dagum's decomposition of the Gini index, alongside the two traditional components of inequality within, G_w, and inequality between the subgroups, G_b, also introduces a component related to overlap, G_o. The differences between the poor and non-poor units according to the environmental dimension are jointly evaluated by means of G_b and G_o.

The inequality within component G_w is simply obtained as a weighted average of the Gini indices of the subgroups:

$$G_w = G_{pe}p_{pe}s_{pe} + G_{npe}p_{npe}s_{npe}$$

Given $\bar{y}_{pe} < \bar{y}_{npe}$, the components of inequality between subgroups G_b and of inequality related to overlap G_o are derived from $G_{pe.npe}$ and $G_{npe.pe}$, attributing to G_b the differences $|y_{pei} - y_{nper}|$ if $y_{pei} < y_{nper}$ and to G_o the differences $|y_{pei} - y_{nper}|$ if $y_{pei} > y_{nper}$. It is also possible to obtain (see Costa [6]) simplified expressions for both G_b and G_o as

$$G_b = p_{pe} - s_{pe} + G_o$$

and

$$G_o = (G - G_w - p_{pe} + s_{pe})/2.$$

from which it is extremely clear the central role that the overlap plays in measuring the differences between the poor and non-poor groups.

In order to take into account the effects of the environmental dimension on the Sen index, it is finally possible to decompose G_p, which allows to further investigate the role of environmental conditions on the income distribution among the poor.

3 A Case Study on Italian Data

With the aim of illustrating the previous methods, we develop a case study on the data from the Bank of Italy's Survey on Households Income and Wealth for 2006, which is, unfortunately, the last year in which information on environmental conditions is available. The survey refers to a representative sample of the Italian population, the size of which allows reliable results at the national and macro territorial level (north-west, north-east, center, south and islands), while, starting from the regional level, the sample size may be insufficient. A complete review of the methodological issues related to the survey is provided in Baffigi et al. [3].

The variable of interest in order to evaluate the environmental dimension classifies the location of dwellings into four groups: degraded areas, neither prestigious nor degraded areas, prestigious areas, other.

In reference to the poverty lines, we set the 60% of the median of the equivalent income as z_y, the poverty line for equivalent income, and we consider households living in degraded areas to be environmentally poor.

Table 2 reports the size of the different groups and also their sample weight, which is the quantity of interest in order to calculate the various indicators, as exemplified for H in Sect. 2.1.

On the basis of z_y, we detect 1314 poor units, while, on the basis of z_e, we have 350 poor units. Table 2 also shows how only half of the poor units according to the environmental dimension are also poor according to income, while 85% of the non-poor according to the environmental dimension are also not poor according to income: a good match is highlighted between the groups of the non-poor, while there are strong differences between the groups of the poor.

3.1 Unidimensional Indicators

The first results which can be derived from Table 2 are the unidimensional head count ratios for the two inequality dimensions which are analyzed here.

As for income, $H_y = 0.181$ indicates the presence of 18.1% of poor families, while $H_{y|ep} = 0.496$ and $H_{y|enp} = 0.165$ show really different dynamics for the conditional distributions, indicating a relevant presence (49.6%) of poor according to income among the poor according to the environmental dimension, against only a 16.5%

Table 2 Italian households 2006, poor and non-poor by income and environmental conditions

		Income		
		Poor	Non poor	
Environment	Poor	$q = 176$	$q_e - q = 174$	$q_e = 350$
		$a_q = 184.39$	$a_{qe-q} = 187.71$	$a_{qe} = 372.10$
	Non poor	$q_y - q = 1138$	$n - q_e - q_y + q = 6280$	$n - q_e = 7418$
		$a_{qy-q} = 1217.78$	$a_{n-qe-qy+q} = 6178.12$	$a_{n-qe} = 7395.90$
		$q_y = 1314$	$n - q_y = 6454$	$n = 7768$
		$a_{qy} = 1402.17$	$a_{n-qy} = 6365.83$	$a_n = 7768$

of poor according to income among the non-poor according to the environmental dimension.

Furthermore, even if only in a descriptive context, the strong difference between $H_{y|ep}$ and $H_{y|enp}$ suggests to reject the hypothesis of perfect overlap.

For the environmental dimension, $H_e = 0.048$ implies 4.8% of poor families, with $H_{e|yp} = 0.131$ and $H_{e|ynp} = 0.029$: the conditional distributions still emphasize the presence of strong differences.

The overall income inequality indicator provided by the Gini index is equal to 0.318, while, conditionally to the income-poor units only, we have $G_{py} = 0.161$, and $G_{npy} = 0.264$ for the income-non-poor, thus suggesting a low inequality level among the poor units according to income.

Furthermore, we can obtain $G_{pe} = 0.314$ on the 350 units that are poor according to the environmental condition and $G_{npe} = 0.328$ on the environmental-non-poor units, indicating a non negligible inequality in both cases.

Finally, the poverty Sen index is equal to 0.071 and it combines the size of poverty, evaluated by $H_y = 0.181$, the depth of poverty (measured by $I_p = 0.279$) and its distribution, summarized by $G_{yp} = 0.161$.

The unidimensional poverty indicators are summarized in Table 3.

Table 3 Results for income and environmental inequalities, unidimensional indicators, Italian households 2006

Head count ratio			
$H_y = 0.181$	$H_{y\|ep} = 0.496$		$H_{y\|enp} = 0.165$
$H_e = 0.048$	$H_{e\|yp} = 0.131$		$H_{e\|ynp} = 0.029$
Gini index			
$1G = 0.318$	$G_{py} = 0.161$		$G_{npy} = 0.264$
	$G_{pe} = 0.328$		$G_{npe} = 0.314$
Sen index components			
$S = 0.071$	$H_y = 0.181$	$I_{py} = 0.279$	$G_{py} = 0.161$

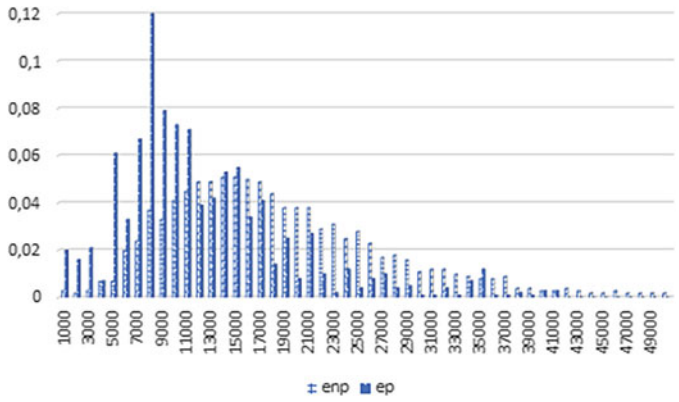


Fig. 1 Poor (ep) and non poor (enp) according to the environmental dimension—Italian households income 2006

3.2 Bidimensional Indicators

The first step of the joint analysis of the two dimensions can be illustrated by means of the picture of the income distribution of the two groups of the q_e poor units and the $(n - q_e)$ non poor units according to the environmental dimension (last column of Table 2).

From Fig. 1 it is possible to observe how the two groups differ in many aspects, showing quite different distributions, but also sharing a strong overlap.

The probability of transvariation, which evaluates the frequency of overlap between the two dimensions, is 50.3%; the extent of overlap is measured by the intensity of transvariation, which is equal to 39%, thus indicating a non-negligible, albeit not high, value.

Conditionally to income-poor units only (first column of Table 2) we have a different overlap between the poor and non poor for the environmental dimension: probability and intensity of transvariation are 0.79 and 0.72, respectively.

Moving to the multidimensional indicator H_{ye}, which takes into account the two dimensions jointly, we have a percentage of poor families equal to 9.6% for $w_i = log(n/q_i)$, while we have $H_{ye} = 0.116$ for $w_i = 0.5$.

The sets of poor units identified by H_y, H_e and H_{ye} coincide only for the 176 units that are poor for both dimensions, while the multidimensional index also classifies as poor some units which are poor according only to one of the two dimensions. More specifically, on the basis of $w_y = 0.74$ and $w_e = 1.32$, the multidimensional index classifies as poor first the 174 poor units according to the environmental dimension and then some of the poor units according to income. On the contrary, in the equally-weighted index, all poor units for only one dimension are considered on the same level, thus inevitably favouring the poor units only for income, which are far more numerous.

Table 4 Results for income and environmental inequalities, multidimensional indicators, Italian households 2006

Two-dimensional indicator					
$w_y = 0.74$	$w_e = 1.32$	$H_{ye} = 0.096$			
$w_y = 0.50$	$w_e = 0.50$	$H_{ye} = 0.114$			
Gini index decomposition					
$G = 0.318$	$G_w = 0.290$	$G_b = 0.023$	$G_o = 0.005$		
$G_{py} = 0.161$	$G_{pyw} = 0.121$	$G_{pyb} = 0.025$	$G_{pyo} = 0.015$		
$G_{pe} = 0.328$	$G_{pew} = 0.104$	$G_{peb} = 0.223$	$G_{peo} = 0$		
Sen index components					
$S = 0.071$	$I_p = 0.279$	$G_{py} = 0.161$	$G_{pyw} = 0.121$	$G_{pyb} = 0.025$	$G_{pyo} = 0.015$

To include the environmental dimension in the analysis, it is also possible to decompose the Gini index on the basis of the two groups of 350 and 7418 units which are, respectively, poor and non-poor with respect to the environmental conditions.

From Table 4, it is possible to observe how the inequality within represents the most relevant component, with a weight of $G_w/G = 91.19\%$, while the inequality between and the overlap component weighs $G_b/G = 7.23\%$ and $G_o/G = 1.57\%$, respectively.

This result should not be read as an indication that the environmental dimension is of little relevance, but as an effect of the weight (94.2%) of the non-poor group on the total. If we look at the decomposition of G_{py}, the Gini index related only to the q_y units which are poor for income, the results are quite different: the inequality within weighs $G_{pyw}/G_{py} = 75.16\%$, the inequality between $G_{pyb}/G_{py} = 15.53\%$ and the overlap component $G_{pyo}/G_{py} = 9.32\%$. We can include these three components in the Sen index computation and evaluate their effect on S.

Furthermore, if we decompose G_{pe}, the Gini index for the environmental-poor units, we are able to assess the differences between the $q = 176$ units poor also for income and the $(q_e - q) = 174$ units poor only for environment. For the decomposition of G_{pe} we obtain, obviously, a zero overlapping, since we compare two groups defined on the basis of the poverty line z_y, and we detect a relevant weight of the inequality between, with $G_{peb}/G_{pe} = 68.1\%$ and $G_{pew}/G_{pe} = 31.9\%$, thus suggesting relevant differences between the two groups.

The summary of the bidimensional poverty indicators is shown in Table 4.

Overall, we identify significantly different patterns for the poor and the non-poor with respect to the environmental dimension, with a strong overlap with the economic dimension only for the non-poor group.

4 Conclusions

Multidimensional poverty indicators and Gini index decomposition allow powerful insights into the interaction between environmental and economic dimensions, especially in reference to the overlap between the two dimensions. These methods are extremely flexible with respect to different definitions of the environmental dimension and can be implemented using a wide range of variables of any type.

In order to assess and to stress the relevance of our results, in a netx paper we aim to develop uncertainty measures of the indicators used in this work, thus allowing to include inferential aspects in the analysis.

As we also find in a case study on Italian data, the poor and non-poor groups show significantly different patterns in the interaction between environmental and economic inequalities.

Environmental poverty is typically stable over time, it is likely linked to persistent poverty, and can therefore be extremely useful in correctly identifying poor units. Our finding confirms and underlines that income alone provides only partial information on poverty conditions, information which can be complemented by the environmental dimension, so as to achieve a relevant improvement both for the poverty measurement and for the implementation of policy actions.

By contributing to the measurement and the evaluation of the interplay between economic and environmental inequalities, we also tackle the interconnession between different SDGS, which represents an essential prerequisite for their successful accomplishment.

References

1. Adger, W.N., de Campos, R.S., Siddiqui, T., Szaboova, L.: Commentary: inequality, precarity and sustainable ecosystems as elements of urban resilience. Urb. Stu. **57**, 1588–1595 (2020)
2. Agusdinata, D.B., Aggarwal, R., Ding, X.: Economic growth, inequality, and environment nexus: using data mining techniques to unravel archetypess of development trajectories. Env. Dev. Sust. **53**, 6234–6258 (2021)
3. Baffigi, A., Cannari, L., D'Alessio, G.: Fifty years of household income and wealth surveys: history, methods and future prospects. Banca d'Italia, Questioni di Economia e Finanza (Occasional Papers) **368** (2016)
4. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In: Dagum, C., Zenga, M. (eds.) Income and Wealth Distribution, Inequality and Poverty, pp. 272–284. Springer, Berlin (1990)
5. Choi, P., Min, I.: Measuring environmental inequality from air pollution and health conditions. Appl. Econ. Lett. **27**, 615–619 (2020)
6. Costa, M.: The Gini index decomposition and the overlapping between population subgroups. In: Mukhopadhyay, N., Pratim Sengupta, P. (eds.) Gini Inequality Index: Methods and Applications, pp. 63–91. Routledge CRC Press (2021)
7. Dagum, C.: A new approach to the decomposition of the Gini income inequality ratio. Empirical Econ. **22**, 515–531 (1997)
8. Downey, L.: Assessing environmental inequality: how the conclusions we draw vay according to the definitions we employ. Sociol. Spectr. (2005). https://doi.org/10.1080/027321790518870

9. Meya, J.N.: Environmental inequality and economic valuation. Environ. Res. Eco. (2020). https://doi.org/10.1007/s10640-020-00423-2
10. Millimet, D.L., Slottje, D.: An environmental Paglin-Gini. Appl. Econ. Lett. **9**, 271–274 (2002)
11. Schaeffer, Y., Tivadar, M.: Measuring environmental inequalities: insights from the residential segregation literature. Ecol. Econom. **164**, 1–14 (2019)
12. Sen, A.K.: Poverty: an ordinal approach to measurement. Econometrica **44**, 219–231 (1976)

The Italian Social Mood on Economy Index During the Covid-19 Crisis

A. Righi, E. Catanese, L. Valentino, and D. Zardetto

Abstract Since 2016, Istat has published the Social Mood on Economy Index (SMEI), an experimental high-frequency sentiment index derived from public tweets in Italian. Since the economic shock produced by the Covid-19 pandemic has not significantly affected the SMEI series, we wondered to what extent the SMEI could grasp the change in the mood due to the pandemic. We produced alternative sentiment indicators, and we compared them to nontraditional high-frequency series to assess the coherence of the SMEI. An alternative index, calculated by introducing pandemic-related terms in the lexicon used for sentiment analysis, better grasped the negative economic trend during the pandemic. We concluded that a continuous adaptation of the dictionary in lexicon-based techniques could improve the coherence.

Keywords Covid-19 · Sentiment analysis · Twitter

1 Introduction

In recent years, the Italian National Institute of statistics has exploited social media messages to assess the Italian mood on the country's economic situation. In October 2018, this effort led to the release of the Social Mood on Economy Index (SMEI)

Authors' contributions—DZ was a major contributor in writing para 3.1, EC in writing para 4.1 and LV in writing para 4.2. AR was a major contributor in writing paragraphs 1, 2, 3.2 and 3.3.

A. Righi (✉)
Istat, Via C. Balbo, 16, 00100 Rome, Italy
e-mail: righi@istat.it

E. Catanese · L. Valentino
Istat, Via C. Balbo, 39, 00100 Rome, Italy
e-mail: catanese@istat.it

L. Valentino
e-mail: luvalent@istat.it

D. Zardetto
World Bank, Via Labicana, 110, 00184 Rome, Italy
e-mail: dzardetto@worldbank.org

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_29

[1], an experimental high-frequency sentiment index based on Twitter data. The new index, derived from samples of public tweets (in Italian) captured in real-time, has gained a good spread among economic analysts.

When the pandemic crisis of Covid-19 began in Italy, In February 2020, the collapse of the main macroeconomic variables and the consumer confidence indicator witnessed the seriousness of the economic crisis continuing till in early 2021 [2]. Since the economic shock has not significantly affected the SMEI series, we wondered to what extent the SMEI was able to grasp the mood change due to the pandemic.

Our work aimed both to present the new statistical tool for economic analysis by studying its relationships with other high-frequency indicators coming from nontraditional sources to understand the reasons for the relative unresponsiveness to the economic shock. We posed two research questions: was the SMEI sufficiently inclusive of the messages on the emergency during the pandemic? Can we calculate alternative sentiment indices to the SMEI capable of better capturing the economic trend distorted by the pandemic crisis?

Thus, we derived from SMEI both the share of tweets containing the terms Coronavirus or Covid and the Covid SMEI (an alternative index calculated by introducing three words related to the pandemic in the lexicon-based approach) and we compared them with Google Trends series.

2 Background

Sentiment analysis is an increasing area of research and application due to the enormous amount of unstructured natural language data currently available. Different studies presented the main algorithms for text mining and sentiment analysis [3, 4]. The sentiment analysis on social media short texts can be performed in different ways. There are methods based on unsupervised learning techniques [5], others on lexicon-based methods [6], and others on combinations of two previous approaches [7]. Some techniques, developed through network measures and clustering techniques, use polarization, controversy, and topic tracking in time [8]. Others use the hashtags classification developed through probabilistic models [9].

Analysis making use of lexicon-based methods mainly refers to the English language [10–12]. There is a lesser development of the lexicon in Italian language. Basile and Nissim [13] developed the Sentiment Italian Lexicon (Sentix), a dictionary whose lemmas are associated with pre-computed sentiment scores. Istat followed the lexicon-base method and used the Sentix lexicon in developing the experimental high-frequency sentiment index on the overall economic situation from Twitter data [14].

Kharde and Sonawane [15] showed it is very effective in cases that require little effort in human-labeled documents. The combination of lexicon-based methods and learning methods may improve both the accuracy and the capacity to adapt to various domains.

3 Data and Methods

3.1 Social Mood on Economy Index

The SMEI is an experimental daily-frequency index that aims to measure the public sentiment on the economy based on messages of the Italian Twitter users. Audiweb [16] quantified in 2019 more than 10.2 million Italian users of the microblogging service created in 2006, which lets users post texts limited in length (tweets). We used Twitter's streaming application programming interface (API) to collect samples of public tweets matching a filter made-up of 60 keywords relevant to the study of the general and personal economic dimensions. We computed daily index values by processing all the filtered tweets reported in a single day as a single block. We processed around 57,000 daily tweets containing at least one of the selected hashtags from February 2016 to the end of June 2021. We pre-processed the tweets through different phases of text cleaning and normalization. We converted every letter of a word into lowercase and tokenized the text into words. We applied basic orthographic repairs and removed uniform resource locators or non-alphabetic characters and stop words. We concluded this process with the lemmatization of words to return the base form of the word, removing inflectional endings. Twitter users often use slang, emoticons, emojis, acronyms, and punctuation to signal feelings more intensely. We supposed that sarcastic tweets have little to no effect on aggregate sentiment metrics, following [17].

We used a mixed approach with lexicon-based methods and unsupervised techniques for sentiment analysis [14]. We clustered the tweets according to their sentiment scores into three mutually exclusive groups (then labeled as Positive, Negative, and Neutral).

The index value was derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes and is linearly transformed.[1] The index is equivalent to the weighted average of intensities. This reduces the day-to-day volatility, which is a relevant aspect of high-frequency indices so that the index becomes more resilient to misclassifications. An automatic outlier detection procedure discriminate truly anomalous data from correct data within the daily time series. We imputed index values classified as truly anomalous via a multivariate interpolation (nearest-neighbor interpolation method). However, different biases can affect estimates; the free download from Twitter API did not allow full access to tweets generated by all the Italian active users and the Italian users are not a representative sample of the Italian population due to different Twitter penetration rates among various sub-population (e.g., young people).

We concentrated on the period from 1 January 2020 to 30 June 2021. As the period is short, the standard econometric analysis would face serious hindrances and possibly turn out unsuitable. We mainly used graphical analysis to compare the SMEI

[1] The transformation makes equal to zero the SMEI long-run mean (referred to the baseline period 10 February 2016–30 September 2018).

and the derived series to other high-frequency series from nontraditional sources. In the comparisons, we used weekly averages to smoothen the signal.

3.2 SMEI Derived Series

We introduced the pandemic dimension both in the texts of the filtered tweets and the dictionary of the corpus. We calculated the series of the share of tweets containing the terms Coronavirus or Covid in the texts (among those used for the SMEI) out of the total tweets used to compute the SMEI. We normalized these SME Covid share series mapping to 1 the spike recorded on 24 February 2020 (when 31% of tweet texts contain the words Coronavirus or Covid).

The incoming neologisms that strongly characterize the new messages may result in a distortion in the sentiment deduction when a lexicon-based approach is used. For example, the sentence "*This Covid cases growth will certainly lead to a new lockdown*" with the SMEI current lexicon is classified as neutral (sentiment scores: 0.19 positive and 0.17 negative). While a similar sentence, such as "*This growth in bronchopneumonia cases will certainly lead to a new closure*" is classified as negative. The difference is due to the keywords Covid and Lockdown, which do not appear in the SMEI current lexicon.

Thus, we developed an alternative index, named Covid SMEI, adding three Covid-19 related terms to the sentiment lexicon (Covid, Coronavirus, and Lockdown). We attributed the sentiment scores of terms such as bronchopneumonia to the terms 'Covid' and 'Coronavirus', and the scores of the term 'Isolation' to the term 'Lockdown'.

3.3 Other Nontraditional High-Frequency Series

Using Google Trends as a data source, we calculated the weighted average of the query shares referring to the keywords Coronavirus and Covid for the Italian territory. We weighted the series with the Google Trends relative popularity levels of the two keywords (15 for 'Coronavirus' and 9 for 'Covid' on average during the period). We normalized our Google Trends weekly series by mapping to 100 its spike in the period (observed on the week of 9 March 2020).

We considered also the weekly series of the number of Covid-19 new positive cases coming from the Civil Protection Department database and we calculated the Covid-19 positivity rates, namely, the number of new positive cases with respect to total Covid-19 tests performed.

4 Main Results

4.1 SMEI and SME Covid Share Series

SMEI showed a sharp increase in the volume of tweets since March 2020. Tweets more than doubled in two days (from around 67 thousand tweets on 3rd March 2020 to 115 thousand ones on 5 March 2020) and reached 144 thousand on 28 March. After the end of May, tweets returned to the same volume as before the Covid-19 crisis. During the later lockdown periods, the total number of tweets of SMEI did not show any increase.

Signals extracted from the weekly SME Covid share series and Google Trends series were almost the same (0.96 of correlation) and showed the same behaviors during the different Covid waves (Fig. 1). While the Google Trends series spiked on the second week of March 2020, when the lockdown of activities was announced, the SME Covid share series peaked when the Italian Government proclaimed the first emergency law (23 February 2020). Both series decreased at the beginning of March 2020 and increased before mid-March. At the beginning of the pandemic, there were more concerns related to the impact of the health emergency on the economy. After March 2020, the interest for the keywords reduced by 0.5, and the decrease continued until June 2020 when both series stayed between 0.1 and 0.2. Both series increased again in the middle weeks of August 2020 and during the second lockdown (in October and November 2020). SME Covid share series was higher than the Google Trends series from the end of March to July 2020. Afterward, the behavior of the two series was similar until April 2021, when the Google Trends series decreased less sharply than the SME Covid share series.

The comparison of the weekly SME Covid share series with the Covid positivity rates series showed an overall correlation of 0.62 (Fig. 2). The Covid positivity

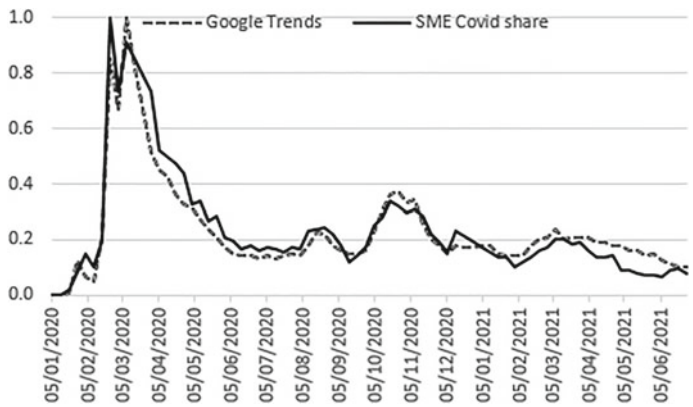


Fig. 1 Comparison of the weekly series of the SME Covid share and the Google Trends Covid/Coronavirus query shares, 1 Jan 2020–30 June 2021

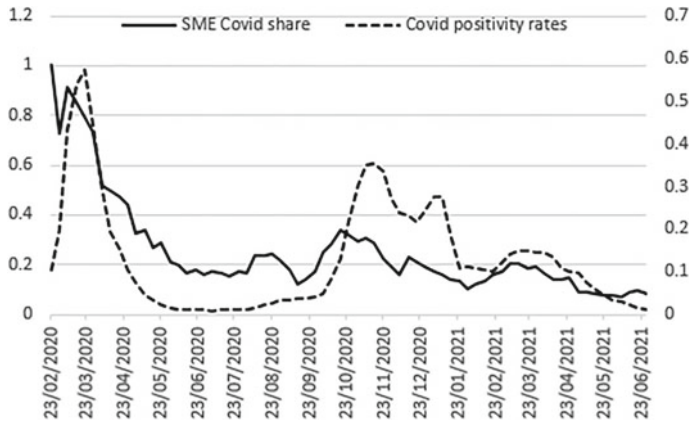


Fig. 2 Comparison between the SME Covid share weekly series and the weekly series of the positivity rates of new Covid cases in Italy, 23 February 2020–30 June 2021

rates spiked on 9 March 2020, when the Government announced the first lockdown. The spikes of the SME Covid share series always anticipated those of the Covid positivity rates during three different Covid-19 waves (March 2020, November 2020, and December 2020–January 2021). It is a possible effect of growing concern about the escalation of the pandemic expressed in the tweets. During the second lockdown, the SME Covid share series peaked on 24 October 2020, anticipating the spike in the Covid positivity rates (9 November 2020). The same happened at the end of 2021, when the SME Covid share series spiked two weeks before the other series (26 December 2020).

The interest in the Covid-19 issue among Twitter users declined with the reduction of the severity of the health emergency from April 2020 to September 2020. SME Covid share series reported almost the same spikes as Covid positivity rates in the three pandemic waves (November 2020, January 2021, and March 2021).

Such similar trends have provided us with proof that the sentiment derived from the texts of the tweets used to calculate the original SMEI provided an undistorted representation of the total speeches that took place on Twitter during the pandemic period. Although, there were no key terms to describe the pandemic threat in our sentiment lexicon.

4.2 The Covid SMEI

We derived our alternative index Covid SMEI adding to the sentiment lexicon of SMEI three Covid-19-related terms, namely ‘Covid’; ‘Coronavirus’; and ‘Lockdown’. We scored these words with the scores of analogous words already in the lexicon. Then, we calculated the index with the same method adopted for the SMEI.

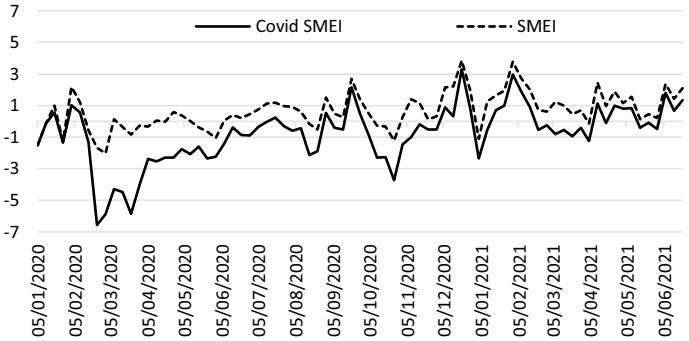


Fig. 3 Comparison between the weekly series of the Covid SMEI and the SMEI, 1 January 2020–30 June 2021

We could assess how the introduction of only three Covid-related terms affected the sentiment reported by the SMEI index analyzing the Covid SMEI. The series of the two indices were highly correlated (0.86). Differences between the series were larger from the weeks after 23 February 2020 to Easter 2020 and from 24 October 2020, to January 2021 (Fig. 3). During these periods, Covid SMEI was largely more negative than the SMEI. Covid SMEI series showed higher volatility especially when the share of tweets containing the term Covid was higher.

A comparison of the weekly moving average differences of SMEI and the Covid SMEI with the Covid SME share series provided us with some insights into the sentiment of tweets related to Covid (Fig. 4). While the SME Covid share series measured the presence of Covid conversation in the current SMEI, the differences between the two indexes give a measure of the impact of the introduction of the Covid terms in the calculation of the sentiment. We can interpret the differences between the two highly correlated (−0.93) series in Fig. 4 as a measure of the relevance of the pessimistic sentiment on the economic situation expressed in the tweets. During the first lockdown (from March to the end of April 2020), the SME Covid share series has declined since its peak at the end of February 2020, while the difference between SMEI and Covid SMEI kept increasing. The difference between SME and SME Covid reached its minimum on the week of March 23, 2020, when the Government closed all non-essential economic activities, a measure with a high economic impact. During the second lockdown (October 2020–January 2021), the difference between SMEI and Covid SMEI was more relevant if compared to the SME Covid share and did not show the spikes reported by the Covid share, keeping more smooth and constant. It could mean that the pessimistic mood on the economic situation affected the Twitter conversation referring to the pandemic a lot.

Moreover, since the positivity rate was a key variable in determining most of the Government policies, we were interested in analysing the behaviour of the Social mood index when the positivity rates were rather high. We wanted to verify which of our series (SMEI and Covid SMEI) was more responsive to changes in the trend of the pandemic cycle. During the first lockdown, when Covid positivity rates

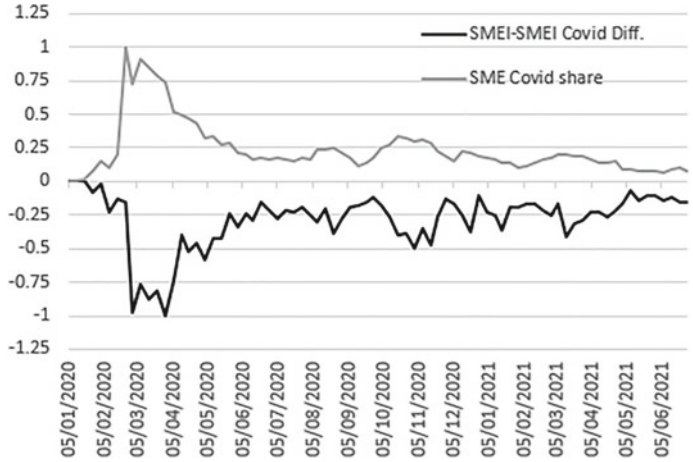


Fig. 4 Differences between the SMEI and the Covid SMEI weekly series compared to the SME Covid share weekly series, 1 January 2020–30 June 2021

peaked, Covid SMEI negatively peaked while SMEI was less negative (Fig. 5). From May 2020 to early October 2020, when positivity rates were low, the Covid SMEI improved, as did most of the economic indicators (GDP). At other times, our series seemed to react to political events, such as when the second lockdown (24 October 2020) was announced or the new government took office (February 2021). In the first case, the Covid SMEI has deteriorated much more than the SMEI; in the second case, both SMEI and Covid SMEI increased, as also other economic indicators did. We, therefore, observed a negative correlation of the positivity rates with the SMEI Covid (−0.48) and weaker with the SMEI (−0.2) and a greater reactivity of the Covid SMEI series to changes in the trend in positivity rates.

Although our series showed high negative correlations with the Google Trends Covid search series (−0.70 for Covid SMEI and −0.43 for SMEI), the comparison showed that the Covid SMEI and, to a lesser extent, the SMEI has had a higher reactivity than the Google Trends series during the political crisis in January 2021 and after the nomination of Mr. Draghi as Prime minister in February 2021 (Fig. 6).

5 Conclusion

Findings showed that SMEI has correctly identified the Covid-related speeches although it proved to be an unresponsive indicator during the pandemic. This implies that the collected tweets gave an undistorted representation of the total conversations that took place.

However, the introduction of some key terms of the pandemic in the dictionary allowed the alternative index to better grasp the negative economic trend recorded by

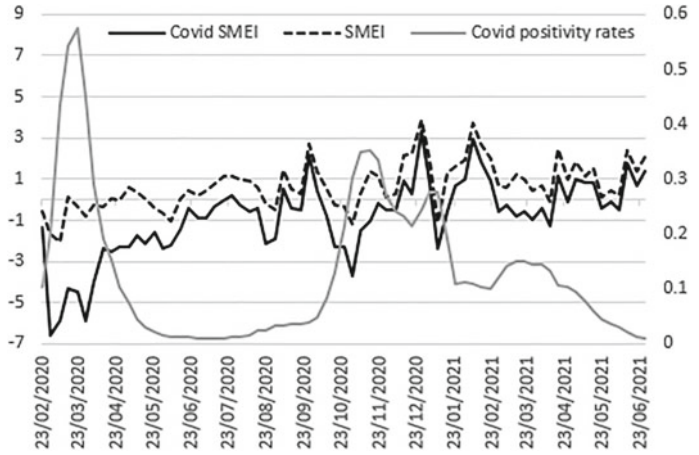


Fig. 5 Comparison between the Covid SMEI, the SMEI, and the positivity rates of new Covid cases in Italy (weekly series), 23 February 2020–30 June 2021

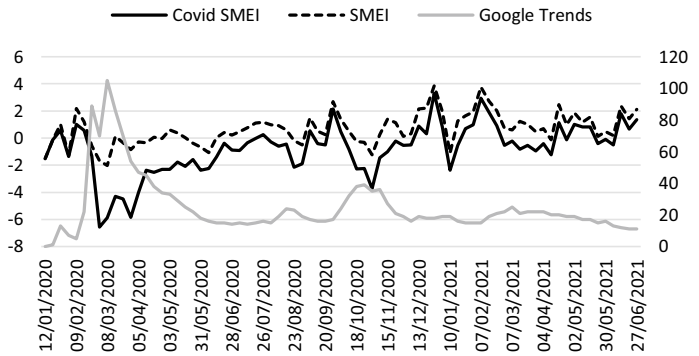


Fig. 6 Comparison between the weekly moving average series of the Covid SMEI, the SMEI, and the Google Trends Covid/Coronavirus query shares), 1 January 2020–30 June 2021

all macroeconomic indicators during the pandemic. During the economic recovery phase, the two indices shortened their differences, although the Covid SMEI showed a sharper improvement in line with all the economic indicators (at least, until March 2021).

Thus, static unsupervised methods for sentiment analysis not allowed capturing the extreme Covid-19 phenomenon, at least in the first phase, when all economic indicators crashed. It is encouraging that introducing only three new words allowed a better interpretation of the dynamics of the economic situation. The negative trend of the Covid SMEI during the economic uncertainty phases showed higher responsiveness compared to the Google Trends Covid series or the Covid SME share series

during all the waves and lockdowns, as a sign of an effective worsening of the economic mood.

We, therefore, derived from the results the decision to modify the SMEI production strategy in order to dynamize our unsupervised lexicon-based techniques. We will achieve this result by introducing new words in the lexicon before evaluating whether to move to supervised dynamic techniques. We believe that a continuous adaptation of the dictionary to lexicon-based techniques could effectively overcome a certain fixity of lexicon-based techniques without having the cost and time-consuming disadvantages of supervised methods.

References

1. Bruno, M., Catanese, E., Iannaccone, R., Righi, A., Scannapieco, M., Testa, P., Valentino, L., Zardetto, D., Zurlo, D.: The Social Mood on Economy Index. Methodological Note. Istat, Rome (2022)
2. Istat: Nota mensile sull'andamento dell'economia italiana 5–6 MAGGIO-GIUGNO 2021 (2021)
3. Feldman, R., Sanger, J., et al.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge U.P., Cambridge (2007)
4. Gandomi, A., Haider, M.: Beyond the hype. Big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 137–144 (2015)
5. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. LREC 2010, 17–23 May 2010, Valletta, Malta, pp. 1320–1326 (2010)
6. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
7. Kolchyna, O., Souza, T.T., Treleaven, P., Aste, T.: Twitter sentiment analysis: lexicon method, machine-learning method, and their combination (2015). arXiv preprint arXiv:1507.00955
8. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Quantifying controversy in social media. In: The 9th ACM International Conference on Web Search and Data Mining WSDM'16, pp. 33–42. San Francisco, California (2016)
9. Coletto, M., Lucchese, C., Orlando, S., Perego, R.: Polarized user and topic tracking in Twitter. In: SIGIR'16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, July, pp. 945–948 (2016)
10. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC, vol. 6, pp. 417–422 (2006)
11. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
12. Strapparava, C., Valitutti, A.: Wordnet effect: an affective extension of wordnet. In: LREC, vol. 4, no. 1083–1086, p. 40 (2004)
13. Basile, V., Nissim, M.: Sentiment analysis on Italian tweets. In: 4th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, WASSA 2013, June 14th, 2013, pp. 100–107. Atlanta, USA (2013)
14. Zardetto, D.: Using twitter data for the social mood on economy index. In: Atti della XIII Conferenza nazionale di statistica, Rome, 4–6 July 2018, ISBN 978-88-458-2016-8, pp. 385–390 (2020)

15. Kharde, V., Sonawane, P.: Sentiment analysis of twitter data: a survey of techniques. Int. J. Comput. Appl. **139**(11), 5–15 (2016)
16. Audiweb: Total digital audience in Italia. https://www.audiweb.it (2020)
17. Freire-Vidal, Y., Graells-Garrido, E.: Characterization of local attitudes toward immigration using social media. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 783–790 (2019)

The Rating of Journals and the Research Outcomes in Statistical Sciences in Italian Universities

Maria Maddalena Barbieri, Francesca Bassi, Antonio Irpino, and Rosanna Verde

Abstract Since before its first introduction in 2012, the list of journals of "A class" for Statistical Sciences has generated a large debate in the scientific community on its formulation. The list is one of the outcomes of the ranking of journals proposed by ANVUR (the Italian National Agency for the Evaluation of Universities and Research Institutes) with the main purpose of calculating the minimum values of the indicators of scientific qualification used in the National Scientific Habilitation (ASN) procedure. The paper aims to analyze the data currently available on the research outputs of the Italian academic statistical community to check if any of the potentially observed changes may be related to the composition of the rating of journals.

Keywords ANVUR · ASN · Journals rating · Scopus ASJC

1 Preliminaries

Starting from 2012, ANVUR (the Italian National Agency for the Evaluation of Universities and Research Institutes) carries out the rating of journals for the scientific fields identified as "not bibliometric," such as *Economics and Statistics* (Area 13). The rating was first intended to calculate the minimum values of the indicators

M. M. Barbieri (✉)
Università Roma Tre, Rome, Italy
e-mail: marilena.barbieri@uniorma3.it

F. Bassi
Università Di Padova, Padua, Italy
e-mail: francesca.bassi@unipd.it

A. Irpino · R. Verde
Università Della Campania Luigi Vanvitelli, Caserta, Italy
e-mail: antonio.irpino@unicampania.it

R. Verde
e-mail: rosanna.verde@unicampania.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_30

of scientific qualification used in the National Scientific Habilitation (ASN) procedure both for candidates and for full professors applying for membership in the National Committees that examine applicants. The ASN is necessary to apply for permanent positions of full and associate professor in Italian Universities. However, the rating has soon been employed for different purposes, e.g., it has become one of the means used by Universities' Departments to evaluate the research outcomes of their members and to define the eligibility criteria to serve in a selection board for competitions for early-stage university researchers. More recently, the use of the rating has also been introduced in the accreditation procedure of Ph.D. programs.

The rating procedure has actually two outputs: the list of journals considered with scientific content and the list of top journals, called "A class."

We point out that the first rating was carried out starting from the list of journals where assistant, associate and full professors working in Italian Universities published in the previous years. The lists were subsequently completed also after consultation with the Scientific Societies. In 2016, a new procedure to classify the top journals in the area of *Economics and Statistics* was introduced, and the "A class" list for the sub-area of Statistical Sciences was consequently amended. This procedure was only based on the choice of a certain number of ASJC (All Science Journal Classification) in the Scopus database referred to the years 1999–2005, followed by the selection of a top percentage of journals for each ASJC. The values of the percentage of picks were specific for different ASJC. Since then, both lists of "scientific" and "A class" journals were periodically modified, using criteria that also changed from time to time. However, the main core of the current lists is still the one issued in 2016. Our analyses refer to the updated version of the list published in January 2021.

It is worth mentioning that, as part of the evaluation of universities' research quality (VQR) exercise, carried out by ANVUR every five years starting from 2004, in each of the first two exercises, the group of evaluation experts (GEV) for the area of Economic and Statistics compiled their own journal list, following different strategies. The resulting lists were different from the concurrent ASN journal rating on both occasions.

Since before its first introduction, the rating of journals for Statistical Sciences has generated a large debate in the scientific community related both to the structure of the resulting lists and to their sensible uses. Proof of the intense and still actual activities on this topic is all the actions undertaken by the Italian Statistical Society, whose tracks may be found on the web page of the Society,[1] and the number of papers written on this subject. Main contributions to the discussion were given by [2–5], along with authors of other articles that appeared in two special issues of *Statistica & Società*, published in 2008 and in 2014. These papers represent a summary of experiences, observations, ideas, thoughts, and considerations following the work of boards of experts and a number of events and activities on this subject, such as meetings, round tables, and opinion surveys. In particular, they contain the most relevant contributions

[1] https://www.sis-statistica.it (under DOCUMENTI, then *Documenti della SIS* or *Valutazione della ricerca nelle scienze statistiche*).

to the discussion on the requirements a journal has to meet for being classified as "scientific" or "A class" and some worries about the possible influence that a rating of journals may have on the choice of a research topic. The rationale is that researchers may decide to disregard topics out of the subject categories associated with the journals' list and direct their efforts to opportunistically identify the argument of their studies based on the journals present in the list. Since the publication of the ratings by ANVUR, all the efforts are directed toward the proposal of additional titles to integrate the list of journals considered with scientific content and the "A class" list. In the latter case, the aim is mainly the expansion of the original set of application fields (essentially limited to the area of Economics and Business) and the addition of journals with bibliometric measures not far from those of the titles in the list but outside of it due to the sharp cut-off used in the algorithm adopted to compose the ratings.

The purpose of this paper is to contribute to the discussion on this topic by analyzing the data regarding the publications in journals by the assistant, associate and full professors working in Statistical Sciences in the Italian Universities to check how the habit of the publication changed during the last ten years, also taking into account the "A class" list of journals and the criteria used to compose and maintain it.

The paper is organized as follows: the next section contains an analysis of the publication in journals by the Italian academic researchers in Statistical Sciences, based on graphical representations and descriptive statistics. Some conclusions at the end.

2 The Output of Research of the Italian Academic Statistical Community in the Last Decade

Our analysis is based on publicly available data. The first set of information is the list of researchers working in Statistical Sciences in the Italian Universities in the time interval 2016–2020. From now on, the population of academic researchers will refer to the whole of different existing positions in the Italian university system i.e., full, associate and assistant professors. Where the last category consists of three different sub-categories.

In fact, the former permanent position of assistant professor was canceled in 2010 (a national law established that no new positions could be opened). Since then, the entry-level positions are of two different types of fixed-term assistant professor (or research assistant) called *ricercatore a tempo determinato di tipo a* (RTD-A) or *di tipo b* (RTD-B), depending on the type of work contract linked to the opened position.

Going back to our data, the list was obtained by merging the researchers' lists on the Italian Universities' books on December 31 of each year from 2015 to 2019, downloaded from the Italian Ministry of University and Research repository.[2] Members who retired and new hires during the observation period were also included.

The population considered may be subdivided into three sub-populations, classifying its members by academic recruitment fields pertaining to Statistical Sciences (i.e., 13/D1 Statistics, 13/D2 Economic Statistics, 13/D3 Demography and Social Statistics). Figure 1 shows the distribution of each subpopulation by academic position in each year of the period considered.

To have information on the research outputs published in journals, we resorted to the Scopus dataset since it is used as the base for the rating of journals for Statistical Sciences carried out by ANVUR. For each serial title in the Scopus database, we considered the number of papers published from 2010 to 2020, having as an author at least one of the members of the previously referred population. To identify each of them, we used the corresponding Scopus author ID, which is an identifier assigned automatically to every author of at least one article in the index of Scopus. The search was unsuccessful in a few cases due to the absence of a Scopus ID. Table 1 contains a first summary of the overall population; researchers in Italian universities are classified by academic recruitment field (within Statistical Sciences) and whether or not with a Scopus ID.

Our dataset shows an increase in the number of research outputs present in the Scopus database over the observation period. Figure 2 depicts the evolution of the number of papers published in journals, expressed per 100 academic researchers, for each academic recruitment field in Statistical Sciences and overall. Part of the increasing positive trend in time and the resulting different level of the yearly counts before and after the year 2016 may be explained to some extent noticing that our dataset does not include the publications of those who retired before December 31st 2015 and that were not co-authored with a member of our population (i.e. the academic researchers in the years 2016–2020). In addition, some of the new entries in the population could have been not yet active in the interval 2010–2015 or at least in the first part of it.

From Fig. 2, we note that the lines corresponding to each academic recruitment field share a common long-range trend, although they also exhibit occasional different behaviors. While the number of papers that appeared on serial titles non classified in "A class" (represented in panel b) exhibits growth in the whole interval, the number of papers published in "A class" journals (represented in panel a) shows a decrease in the years 2017 and 2018, followed by a quick recovery. This behavior may be due to the exclusion, announced in 2016, of a large number of journals from the "A class" list, starting from the issues published in 2018. Some of those journals were quite popular until then. In view of these exclusions, the submissions by Italian authors may have moved to other journals.

In order to check if the increasing trend in the number of publications is common to all areas of research or if some research fields contributed more than others to the

[2] https://cercauniversita.cineca.it/php5/docenti/cerca.php.

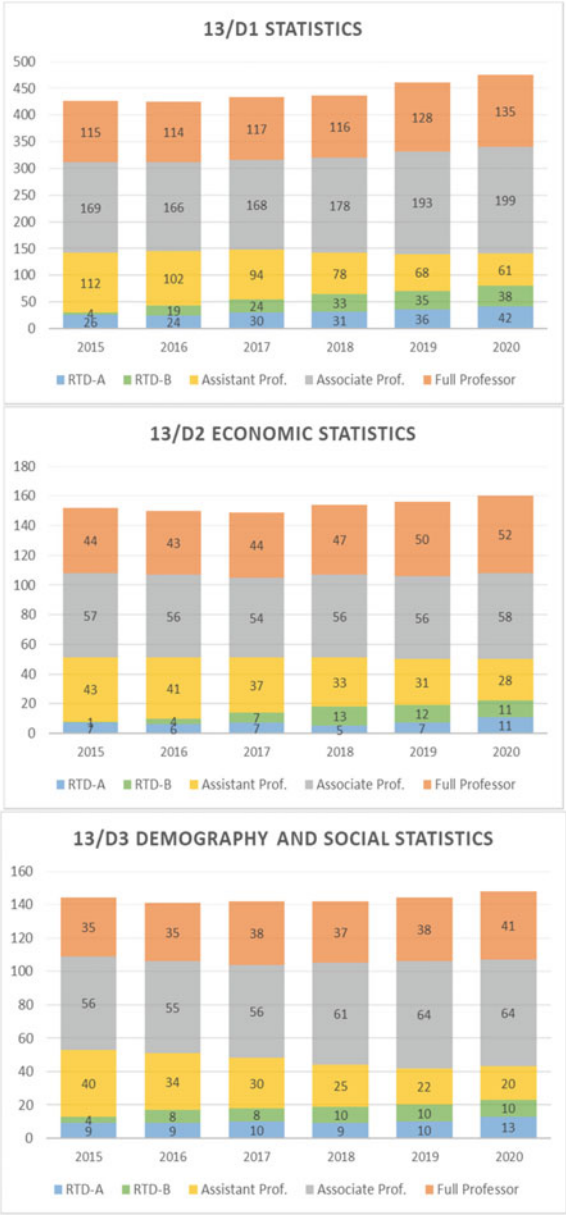


Fig. 1 Distribution of Italian Universities' researchers by academic role in the years 2015–2020

Table 1 Distribution of Italian universities' researchers in the years 2016–2020 by academic recruitment field (*Settore Concorsuale*—SC) and the availability of a Scopus ID

SC	With Scopus ID	Without Scopus ID	Total
13/D1—Statistics	519 (98.3%)	9 (1.7%)	528 (100.0%)
13/D2—Economic Statistics	173 (93.5%)	12 (6.5%)	185 (100.0%)
13/D3—Demography and Social Statistics	162 (93.6%)	11 (6.4%)	173 (100.0%)
Total	854 (96.4%)	32 (3.6%)	886 (100.0%)

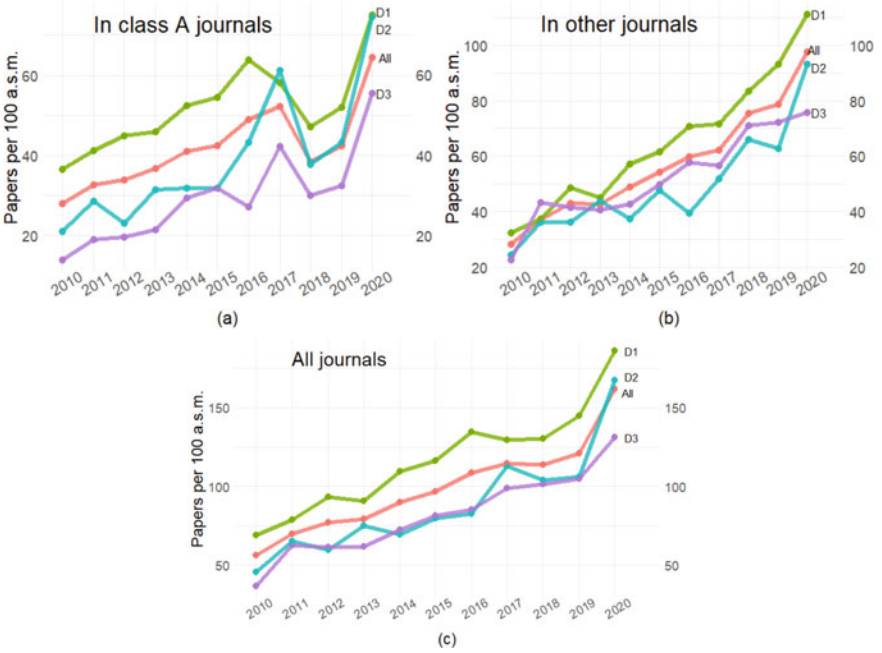


Fig. 2 Number of papers published per 100 academic researchers: panel **a–c** refer to papers published in "A class" journals, in journals that are not in "A class" and in the whole Scopus database, respectively. Please note that the vertical scales are different. (Colored lines key: green—13/D1 Statistics, turquoise—13/D2 Economic Statistics, purple—13/D3 Demography and Social Statistics, red—overall)

changes, we refer to the classification of journals by ASJCs. The ASJC scheme is a journal classification produced by Scopus, based on the aims and scope of the journals and on the content they publish. The classification is not bijective in the sense that each journal may be associated with more than one ASJC. We resorted to the Scopus ASJCs since this classification is the base for the procedure fine-tuned by the working

group appointed by ANVUR to compose the "A class" list of top journals in the area of *Economics and Statistics* [1]. The starting point of this procedure was the choice of a certain number of ASJCs in the Scopus database referred to the years 1999–2005, followed by the selection of a top percentage of journals specific for each ASJC. The group in charge of compiling the "A class" list agreed upon a core of ASJCs common to all academic disciplines in the area of *Economics and Statistics* and a set of specific ASJCs for each sub-area. Thus Statistical Sciences share this common set of ASJCs with other academic disciplines, such as Economics, Economic Policy, Public Economics, Economic History, Business Administration and Accounting Studies, Management, Organization and Human Resources Management, and Commodity Sciences. Most of the serial titles in the "A class" list coming from these ASJCs typically hold papers whose content is usually considered proper of other academic disciplines in the area and not of Statistical Sciences. Also Scopus classifies them as pertaining to different sub-area. This is one of the arguments raised in the debate about the composition of the "A class list" for the potential resulting distortion of the production of the Italian academic statistical community.

As a consequence of the approach adopted by ANVUR to carry out the rating of journals, all journals in the "A class" list are inside at least one of the selected ASJC, apart from a small number of exceptions represented by journals added to the list during the periodical revisions. The set of ASJCs used to compile the A class journal list was our starting point.

For each academic recruitment field, we selected the top ten elements after sorting the ASJCs into decreasing order of the total number of papers published in journals classified in each ASJC in the time period 2016–2020 (denoted as $T_{16–20}$). For each ASJC, we also considered the number of papers published in journals that are not in "A class" (denoted as $NA_{16–20}$), the number of papers published in "A class" journals (denoted as $A_{16–20}$, where $A_{16–20} = T_{16–20} - NA_{16–20}$), the number of "A class" journals involved (denoted as $N_{16–20}$). Then, we computed the difference in the number of papers published in the time intervals 2016–2020 and 2010–2015 with respect to all journals, journals that are not in "A class" and "A class" journals. We denoted the resulting differences as ΔT, ΔA and ΔNA, respectively. Results are reported in Table 2 for the academic recruitment field 13/D1 (Statistics), in Table 3 for the academic recruitment field 13/D2 (Economic Statistics), and in Table 4 for the academic recruitment field 13/D3 (Demography and Social Statistics). The tables also contain the percentages over T, below the number of products classified as A or NA, and the percentage changes below the changes over the two time periods, all in round brackets.

Regarding the set of the most popular ASJCs, although the academic recruitment fields share a common core, they also show differences in the composition of the aggregates and in the proportion of papers in "A class" journals in each ASJC, with 13/D1 usually showing larger values.

While the order of magnitude of frequencies referred to 13/D2 and 13/D3 is in most cases too small to allow comparisons, in 13/D1 data seem to show a change of interest towards research fields different from those usually considered as traditional: the ASJCs with the biggest relative changes in the overall number of papers published

Table 2 13/D1 (Statistics)—Top ten ASJCs with respect to the total number of papers published in 2016–2020

ASJC	$T_{16\text{-}20}$	$NA_{16\text{-}20}$	$A_{16\text{-}20}$	$N_{16\text{-}20}$	ΔT	ΔNA	ΔA
Statistics and Probability	1,330 (100%)	406 (31%)	924 (69%)	66	14 (1%)	2 (0%)	12 (1%)
Statistics, Prob. and Uncertainty	831 (100%)	202 (24%)	629 (76%)	44	84 (11%)	39 (24%)	45 (8%)
Applied Mathematics	356 (100%)	75 (21%)	281 (79%)	26	−3 (−1%)	16 (27%)	−19 (−6%)
Social Sciences (Miscellaneous)	256 (100%)	65 (25%)	191 (75%)	14	135 (112%)	51 (364%)	84 (79%)
Economics and Econometrics	211 (100%)	97 (46%)	114 (54%)	33	103 (95%)	61 (169%)	42 (58%)
Modelling and Simulation	207 (100%)	87 (42%)	120 (58%)	15	2 (1%)	−9 (−9%)	11 (10%)
Computer Science Applications	170 (100%)	72 (42%)	98 (58%)	13	65 (62%)	56 (350%)	9 (10%)
Manag. Sc. and Oper. Research	133 (100%)	32 (24%)	101 (76%)	17	69 (108%)	9 (39%)	60 (146%)
Multidisciplinary	108 (100%)	42 (39%)	66 (61%)	5	48 (80%)	38 (950%)	10 (18%)
Geography, Plann. and Development	105 (100%)	61 (58%)	44 (42%)	12	71 (209%)	38 (165%)	33 (300%)

are *Social Sciences (miscellaneous)*, *Economics and Econometrics*, *Management Science and Operations Research* and *Geography, Planning and Development*. The latter has one of the largest relative changes in 13/D2 and 13/D3 as well. We also note that *Social Sciences (miscellaneous)* is the common leader in the increase of papers published in "A class" journals.

In order to gather information also on the number of papers published in journals classified in more than one ASJC, we refer to the Venn-Euler diagrams reported in Figs. 3 and 4 for the academic recruitment field 13/D1 Statistics, in Figs. 5 and 6 for 13/D2 Economic Statistics, in Figs. 7 and 8 for 13/D3 Demography and Social Statistics, respectively. To plot diagrams, we considered the ten most relevant ASJCs for each academic recruitment field (listed in Tables 2, 3 and 4), while the remaining ASJCs were grouped according to the Scopus area they pertain. In each case, the classification refers to the majority of research products considered. Residual papers are sparse in different Scopus areas, each having a negligible weight. For the meaning of labels of Scopus areas used in the plots, we refer to Table 5 in the Appendix. In Figs. 3, 4, 5, 6, 7 and 8, the area of each bubble is proportional to the number of papers published in journals classified in the corresponding ASJC or Scopus area, while the area of the overlapping regions is approximately proportional to the number of papers in the corresponding ASJCs/Scopus areas have in common. In the case of

Table 3 13/D2 (Economic Statistics)—Top ten ASJC with respect to the total number of papers published in 2016–2020

ASJC	$T_{16\text{-}20}$	$NA_{16\text{-}20}$	$A_{16\text{-}20}$	$N_{16\text{-}20}$	ΔT	ΔNA	ΔA
Economics and Econometrics	248 (100%)	121 (49%)	127 (51%)	40	26 (12%)	17 (16%)	9 (8%)
Statistics and Probability	180 (100%)	81 (45%)	99 (55%)	31	44 (32%)	30 (59%)	14 (16%)
Social Sciences (miscellaneous)	159 (100%)	56 (35%)	103 (65%)	13	69 (77%)	20 (56%)	49 (91%)
Geography, Planning and Development	144 (100%)	90 (62%)	54 (38%)	14	90 (167%)	52 (137%)	38 (238%)
Statistics, Probability and Uncertainty	116 (100%)	30 (26%)	86 (74%)	25	40 (53%)	5 (20%)	35 (69%)
Manag., Monitoring, Policy and Law	80 (100%)	52 (65%)	28 (35%)	8	51 (176%)	36 (225%)	15 (115%)
Strategy and Management	67 (100%)	23 (34%)	44 (66%)	13	33 (97%)	0 (0%)	33 (300%)
Econ., Econometrics and Finance (misc.)	64 (100%)	35 (55%)	29 (45%)	11	7 (12%)	−3 (−8%)	10 (53%)
Finance	54 (100%)	27 (50%)	27 (50%)	10	−5 (−8%)	−3 (-10%)	−2 (−7%)
Business and International Management	45 (100%)	24 (53%)	21 (47%)	6	11 (32%)	8 (50%)	3 (17%)

journals classified in the "A class," the relevance of each category is obviously related to the number of titles included in the list.

Figure 3 refers to papers having at least one author in the academic recruitment field 13/D1 (Statistics), published in journals and classified by ASJC/Scopus area. Venn-Euler diagrams are based on 2,547 products (out of a total of 2,873) printed in the period 2010–2015 and on 3,250 (out of a total of 3,783) printed in the period 2016–2020. Figure 3 shows some interesting patterns. The number of papers published in journals classified in the Scopus area MEDI (Medicine) is almost the same when passing from the first to the second time interval. However, in the latter, the subset of journals also classified in ASJCs with "Statistics" in their name is more relevant. In the period 2010–2015, Scopus areas MATH (Mathematics) and COMP (Computer Science) present an intersection with BIOC (Biochemistry) and MEDI, while in the period 2016–2020 there is a "chain" with ENGI (Engineering) and AGRI (Agricultural and Biological Sciences). The Scopus area BIOC was between Statistics and MEDI, in the first period, while in the second one, it is completely apart. In the period 2010–2015 the ASJC Social Sciences was related to Statistics and Probability, while in the period 2016–2020 the intersection between the two is empty.

The pictures in Fig. 4 still refer to the field 13/D1 (Statistics), but they have been drawn using only the subset of papers published in "A class" journals. The total

Table 4 13/D3 (Demography and Social Statistics)—Top ten ASJC with respect to the total number of papers published in 2016–202

ASJC	$T_{16\text{-}20}$	$NA_{16\text{-}20}$	$A_{16\text{-}20}$	$N_{16\text{-}20}$	ΔT	ΔNA	ΔA
Demography	158 (100%)	73 (46%)	85 (54%)	12	24 (18%)	16 (28%)	8 (10%)
Social Sciences (miscellaneous)	156 (100%)	48 (31%)	108 (69%)	10	87 (126%)	23 (92%)	64 (145%)
Statistics and Probability	87 (100%)	42 (48%)	45 (52%)	14	18 (26%)	16 (62%)	2 (5%)
Geography, Planning and Development	72 (100%)	34 (47%)	38 (53%)	10	50 (227%)	18 (113%)	32 (533%)
Public H., Envir. and Occupational Health	59 (100%)	50 (85%)	9 (15%)	4	2 (4%)	18 (56%)	−16 (-64%)
Statistics, Probability and Uncertainty	55 (100%)	19 (35%)	36 (65%)	12	23 (72%)	10 (111%)	13 (57%)
Economics and Econometrics	53 (100%)	26 (49%)	27 (51%)	12	17 (47%)	7 (37%)	10 (59%)
Strategy and Management	32 (100%)	10 (31%)	22 (69%)	5	25 (357%)	5 (100%)	20 (1000%)
Multidisciplinary	31 (100%)	12 (39%)	19 (61%)	4	17 (121%)	11 (1100%)	6 (46%)
Health (social science)	31 (100%)	21 (68%)	10 (32%)	4	11 (55%)	12 (133%)	−1 (-9%)

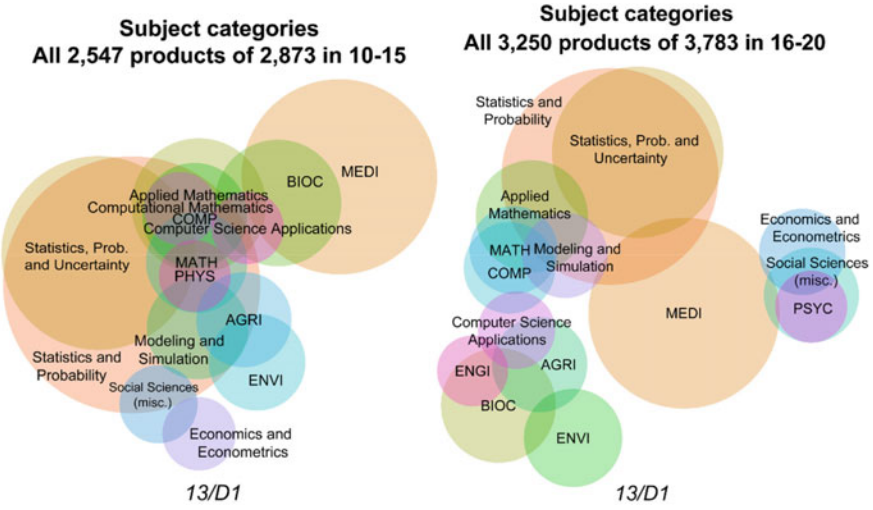


Fig. 3 Venn-Euler diagrams of the papers published in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D1 (Statistics), classified by ASJC/Scopus area

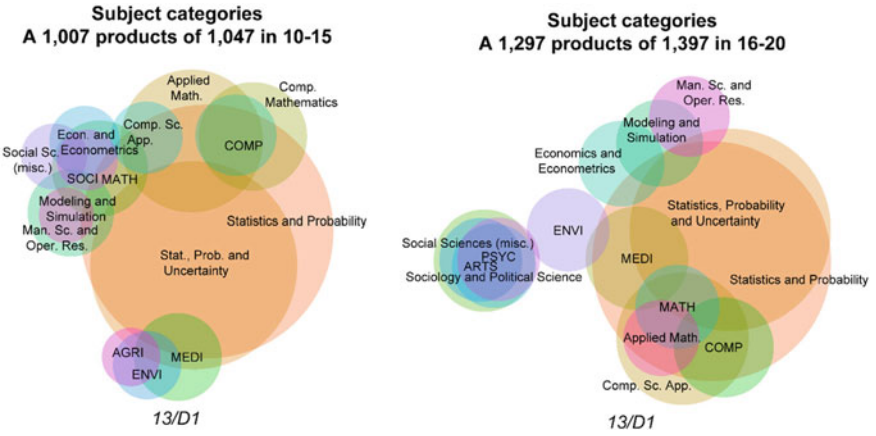


Fig. 4 Venn-Euler diagrams of the papers published in "A class" journals in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D1 (Statistics), classified by ASJC/Scopus area

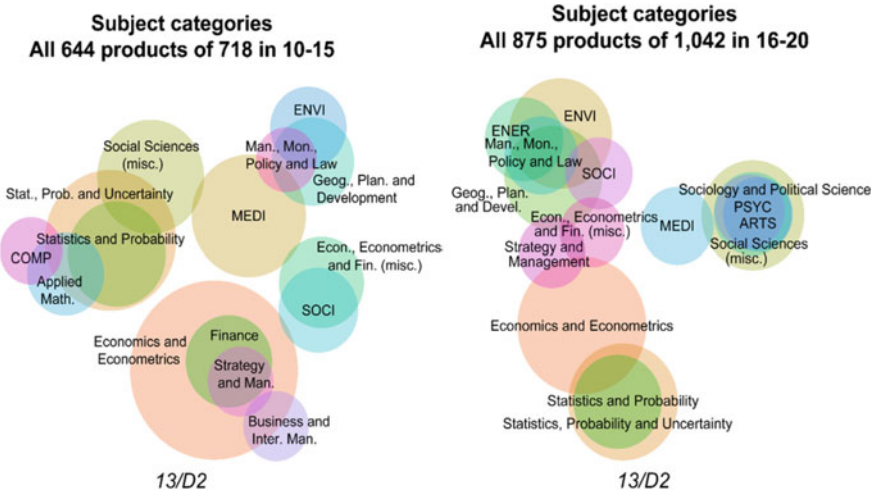


Fig. 5 Venn-Euler diagrams of the papers published in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D2 (Economic Statistics), classified by ASJC/Scopus area

number of papers passed from 1,047 in the period 2010–15 to 1,397 in the period 2016–20. The Venn-Euler diagrams (realized on 1,007 products for the period 2010–15 and on 1,297 for the period 2016–20) point out some differences between the two time intervals. One of the main dissimilarities consists in the reduction of the part the ASJCs more relevant for the field 13/D1 (i.e., Statistics and Probability and Statistics, Probability and Uncertainty) have in common with each of the following

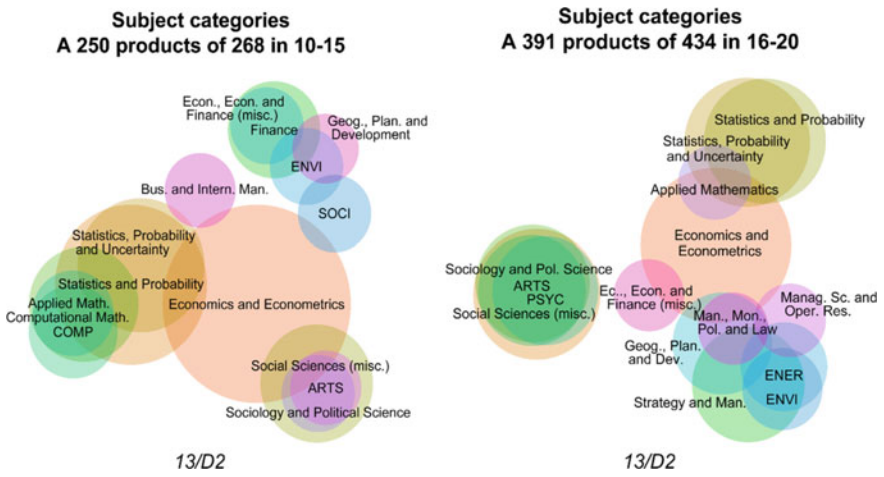


Fig. 6 Venn-Euler diagrams of the papers published in "A class" journals in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D2 (Economic Statistics), classified by ASJs/Scopus area

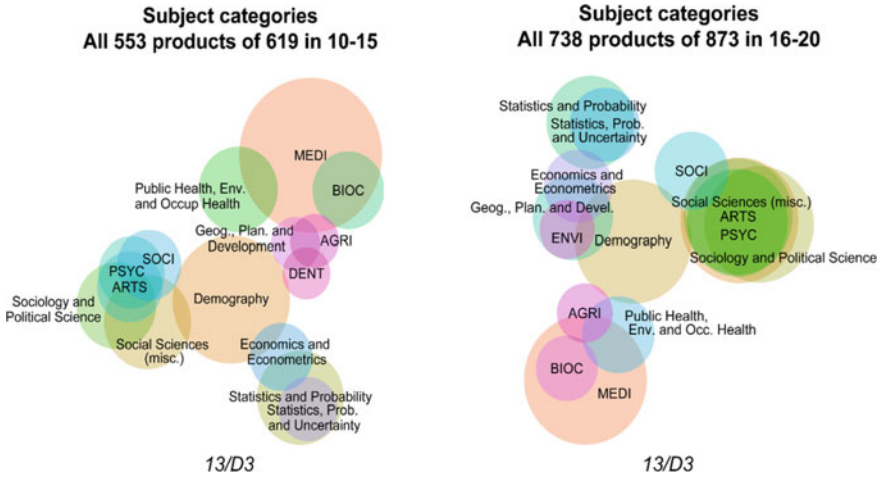


Fig. 7 Venn-Euler diagrams of the papers published in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D3 (Demography and Social Statistics), classified by ASJC/Scopus area

ASJCs/areas: MEDI, MATH, Applied Math, Comp. Sc. App. (Computer Science and Applications). In fact, in the last time interval, the bubbles referred to each of the latter ASJCs/areas are almost completely inside those of Statistics and Probability or of Statistics, Probability and Uncertainty. While in the previous period a considerable part was not in common. This is clearly an effect of the already mentioned exclusion

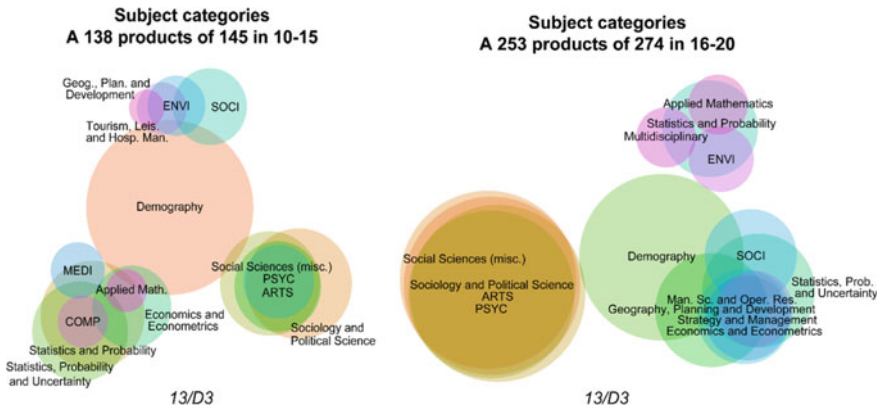


Fig. 8 Venn-Euler diagrams of the papers published in "A class" journals in the years from 2010 to 2015 (left panel) and from 2016 to 2020 (right panel) with at least one author in the academic recruitment field 13/D3 (Demography and Social Statistics), classified by ASJC/Scopus area

from the "A class" list of all journals not classified inside ASJCs considered as pertaining to Statistical Science by the group in charge of carrying out the rating. Another effect of the new formulation of the "A class" list is the increased relevance, in the years 2016–20 over the previous period, of ENVI (Environmental Science), Social Sciences, Economics and Econometrics, mainly involving journals not classified also in Statistics and Probability or Statistics, Probability and Uncertainty. This effect is due to the inclusion in the "A class" list of a number of additional journals in common with the other fields in the area of *Economics and Statistics.* A common feature of both time intervals is the strong overlapping of the Scopus area COMP with the ASJCs more relevant for the field 13/D1.

Figure 5 illustrates the distribution by ASJC/Scopus area of the papers published in journals with at least an author in the academic recruitment field 13/D2. The total number of papers passed from 718 in the period 2010–15 to 1,042 in the period 2016–20. The Venn-Euler diagrams (based on 644 products for the period 2010–15 and on 875 for the period 2016–20) show very different situations in the two time intervals. In the first, there is a group of ASJCs, more related to the field 13/D2, composed of Economics and Econometrics, Finance, Strategy and Management and Business and Inter. Man. (Business and International Management). The first ASJC contains the following two and most of the last one. This group is apart from the rest of ASJCs/Scopus areas, including Statistics and Probability and Statistics Probability and Uncertainty, which appear quite relevant in terms of the number of papers, just as MEDI and Social Science (misc.). In the second time interval, the distribution of the papers appears different. In particular, there are overlaps between Economics and Econometrics and Statistics and Probability and Statistics Probability and Uncertainty on one side, and Econ. Econometrics and Fin (misc.) (miscellaneous) and Strategy and Management, on the other side. This is clearly due to the presence of journals common to those ASJCs. Still apart there are PSYC (Psychology), ARTS,

Sociology and Political Science, Social Science (misc.) and ENVI, which share the feature of presenting an increase over time in the number of products.

Figure 6 shows the distribution of papers published in "A class" journals by ASJC/Scopus area for the academic recruitment field 13/D2. The number of papers passed from 268 in the period 2010–15 to 434 in the period 2016–20. The Venn-Euler diagrams (drawn on 250 products published in the period 2010–15 and on 391 in the period 2016–20) highlight a change in the pattern between the first and the second time interval. The change is mainly characterized by an increase in the number of products published in journals classified in ASJCs different from Economics and Econometrics, Statistics and Probability and Statistics, Probability and Uncertainty. In fact, we observe an increase in the products published in "A class" journals classified in Sociology and Political Science, ARTS, PSY, Social Sciences (misc.), ENER (Energy), ENVI and Strategy and Man. (Management). Bubbles associated with all these categories present only minor overlaps with bubbles associated with ASJCs, which were previously more typical of the field, such as Economics and Econometrics or Statistics and Probability and Statistics, Probability and Uncertainty.

Figure 7 shows the classification by ASJCs/Scopus areas of the papers published in journals, having at least one author in the academic recruitment field 13/D3. The number of papers was 619 in the years 2010–15 and 873 in the years 2016–20. The Venn-Euler diagrams related to the two periods are elaborated on 533 and 730 products, respectively. In this case, there are no significant differences between the two time intervals. In both of them, Demography, Social Sciences, Sociology and Political Science have the largest size in terms of the number of products. The Scopus area MEDI is relevant as well. We also observe an increase over time in the number of publications in journals classified as ARTS, PSYC, ENVI.

Figure 8 contains the classification by ASJC/Scopus area of the papers published in "A class" journals by the researchers in the academic recruitment field 13/D3. The number of papers passed from 619 in the period 2010–15 to 873 in the period 2016–20. The Venn-Euler diagrams are based on 533 products referred to 2010–15 and on 730 referred to 2016–20. We notice the stability in the size of Demography and the considerable increase over time in the number of papers published in "A class" journals classified in two groups of ASJCs. The first group consists of a set of ASJCs whose content is strictly related to the field, i.e., Social Sciences (misc. and Sociology and Political Sciences. While in the second group, we find Man. Sc. (Management Science) and Oper. Res. (Operational Research), Geography, Planning and Development, Strategy and Management and Economics and Econometrics.

In order to study the variability of the distribution of the number of papers published in journals with regard to the top ten ASJCs for each academic recruitment field and for both periods, we computed the Gini concentration index (whose values are reported in Figs. 9, 10 and 11) and drew the box-plots (reported in Figs. 12, 13 and 14).

The Gini concentration index allows us to highlight the presence of journals hosting a number of papers sensibly higher than the others. Except for a few situations, we observe a general increase in concentration over time. For all academic

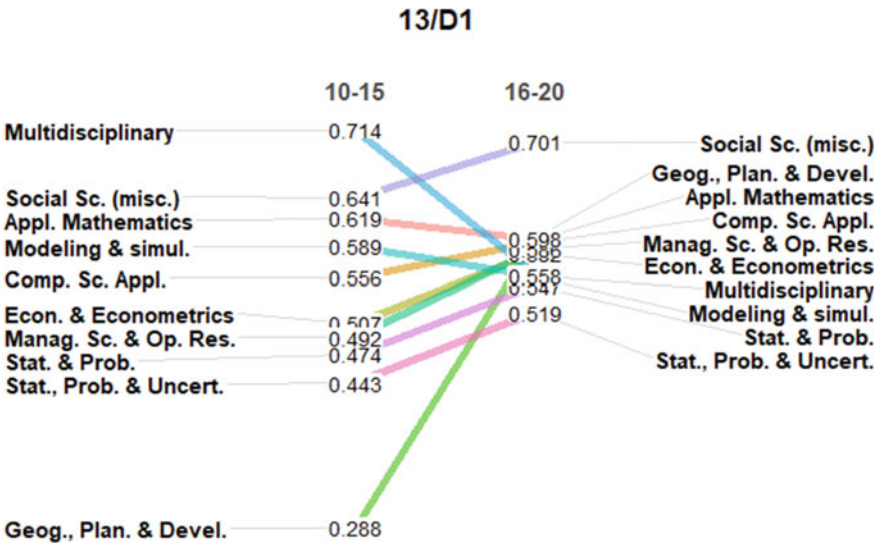


Fig. 9 Variation through time in the Gini index computed on the number of papers published in journals classified in the ten more relevant ASJCs for the academic recruitment field 13/D1

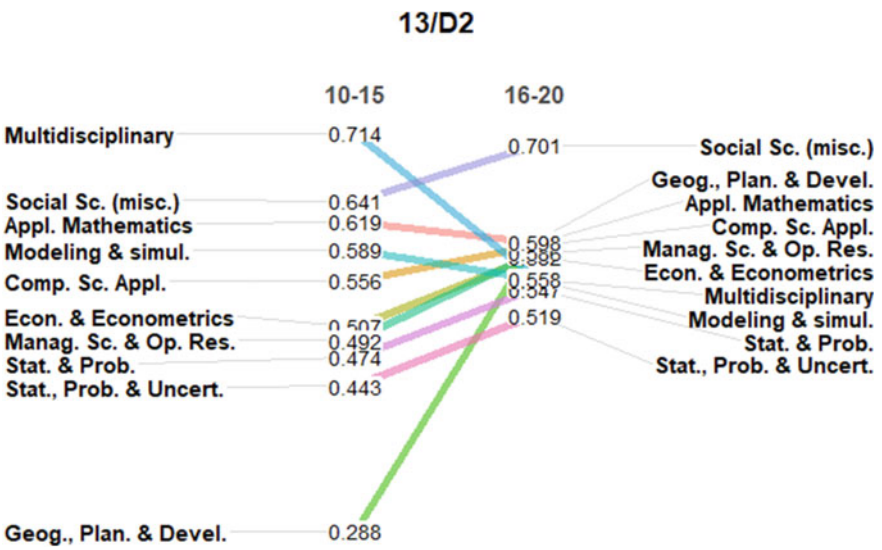


Fig. 10 Variation through time in the Gini index, computed on the number of papers published in journals classified in the ten more relevant ASJCs for the academic recruitment field 13/D2

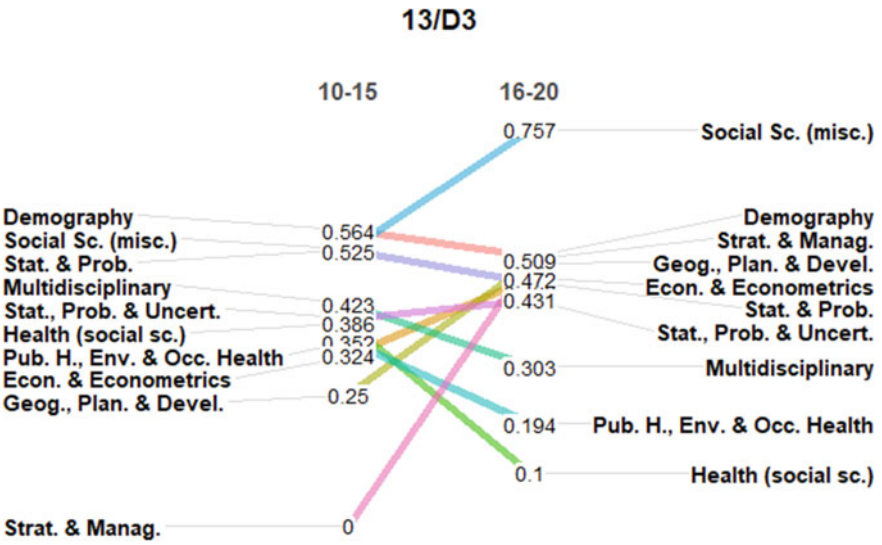


Fig. 11 Variation through time in the Gini index, computed on the number of papers published in journals classified in the ten more relevant ASJCs for the academic recruitment field 13/D3

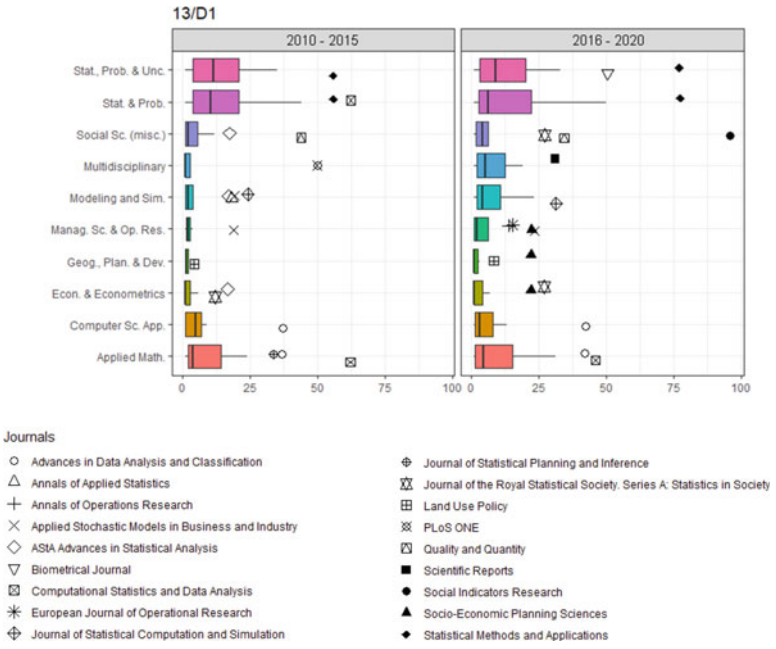


Fig. 12 Box-plots of the distribution of journals by the number of papers published in the ten more relevant ASJCs for the academic recruitment field 13/D1 in periods 2010–2015 and 2016–2020. The observations marked as outliers are larger than the value of the third quartile multiplied by 1.5 times the interquartile range

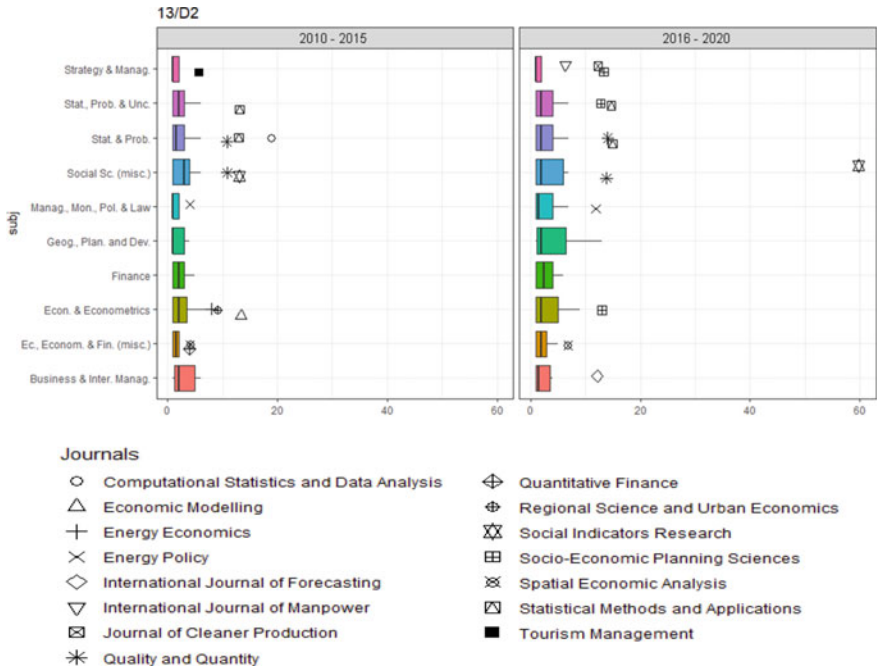


Fig. 13 Box-plots of the distribution of journals by the number of papers published in the ten more relevant ASJCs for the academic recruitment field 13/D2 in periods 2010–2015 and 2016–2020. The observations marked as outliers are larger than the value of the third quartile multiplied by one and a half times the interquartile range

recruitment fields, Social Science is the ASJC with the highest concentration in recent years.

The box plots in Figs. 12, 13 and 14 highlight outlier observations, which contributed to the increase in the value of the Gini index. We note that some outliers are usual publication settings for the Italian academic statisticians. Examples are *Advances in Data Analysis and Classification*, *Biometrical Journal*, *Computational Statistics and Data Analysis*, *Demographic Research, Journal of Statistical Computation and Simulation, Statistical Methods and Applications.* On the other hand, we find journals whose popularity greatly increased only in recent years. This seems to be the case of *Social Indicators Research.*

To analyze the changes in the number of papers published in each journal over time, we resorted to streamgraph plots. The streamgraph plot is an evolution of the stacked area chart for time series, particularly useful to represent the flow of a large number of positive ratio variables or time-varying counts. It differs from the stacked area chart because it has a vertical centered layout with the appearance of a river with different strata. Time series are rearranged and partially smoothed to give a more pleasant aspect. A streamgraph is conceived as an interactive plot. However, for the sake of the paper media, we present a static version of it after annotating the

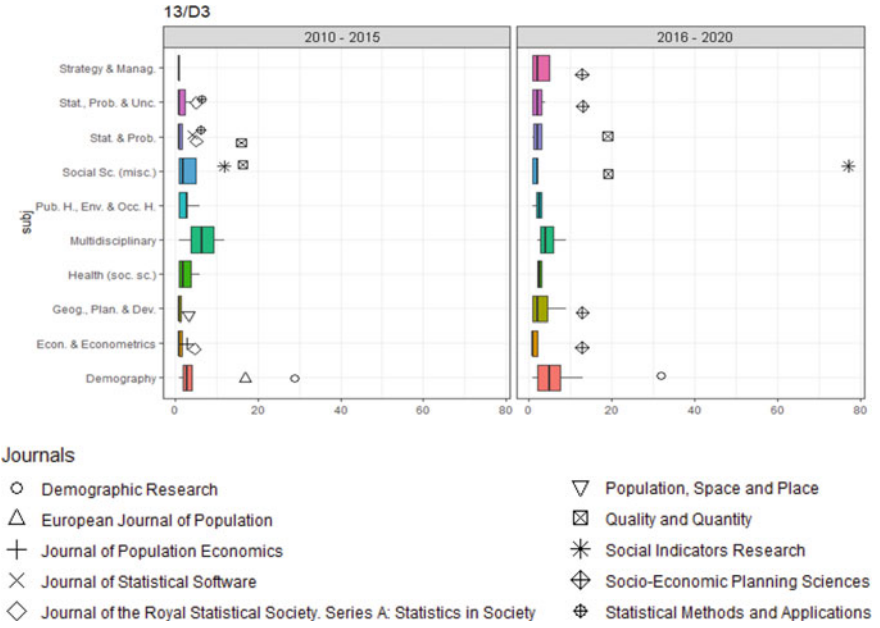


Fig. 14 Box-plots of the distribution of journals by the number of papers published in the ten more relevant ASJCs for the academic recruitment field 13/D3 in periods 2010–2015 and 2016–2020. The observations marked as outliers are larger than the value of the third quartile multiplied by one and a half times the interquartile range

main flows with the name of the corresponding journals. Since the number of journals is very high, for each academic recruitment field, we present two streamgraphs: one for the "A class" journals and one for the others (see Figs. 15, 16, 17, 18, 19 and 20). We remind that after the introduction of the first list of "A class" journals in 2012, a general update of the list was made in 2016 with the announcement of the exclusion of a large number of journals, starting from the issues published in 2018.

Looking at the evolution of the number of papers produced by academic statisticians from 2010 to 2020, we can observe how the modification in the composition of the list has influenced the total flow of "A class" papers.

From our experience, the average time between the submission and the final acceptance of a paper is about one year. This is the reason why in Fig. 15 we can note a significant reduction in the number of "A class" papers published not only in the year 2017 but also in 2018 and 2019. In the meantime, the flow of papers outside the A class list has a continuous growth (Fig. 16). Comparing the two streamgraphs, we can see how the flow of some "A class" journals, e.g., *Plos ONE* and *Quality and Quantity* (which came out from the "A class" list), stopped in Fig. 15 in 2017 and started again in Fig. 16 in the following year.

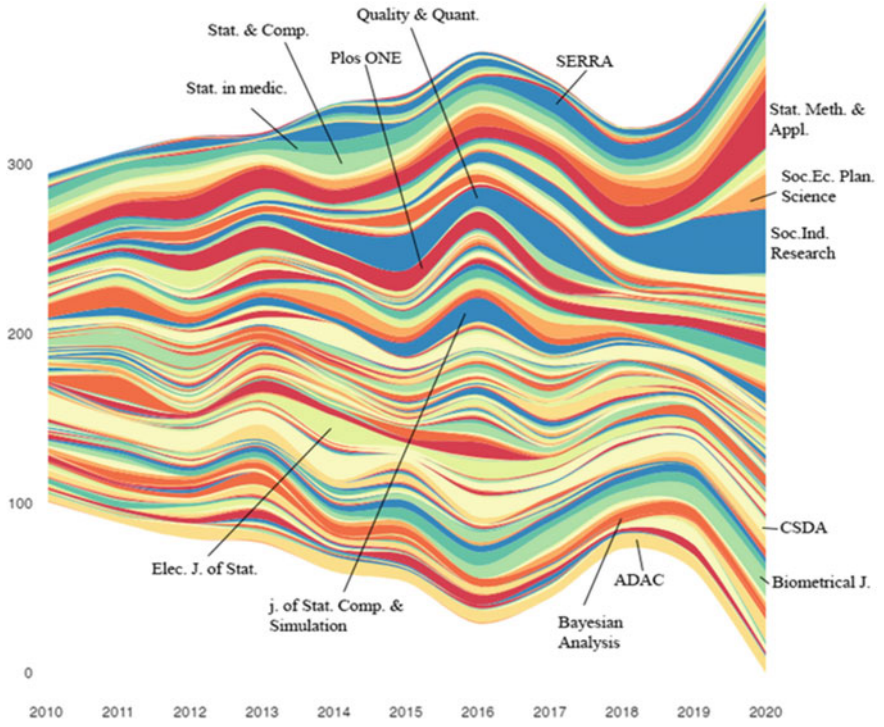


Fig. 15 Streamgraph plot of the number of papers published in "A class" journals for the academic recruitment field 13/D1 (Statistics)

Considering only "A class" journals, we notice that the amount of papers published in a number of journals is almost constant over time. This is the case of *Computational Statistics and Data Analysis*, *Advances in Data Analysis and Classification*, *Biometrical Journal*, and *Bayesian Analysis*, to mention some of them. On the other hand, *Social Indicators Research* and *Statistical Methods and Application* have attracted a number of papers which is increasing over time, especially during the 2018–2020 period. A consistent flow has started from 2016 for *Stochastic Environmental Research and Risk Assessment* (SERRA), a journal related to the applications of statistics to the environmental sciences, while a decreasing production has been observed for *Statistics in Medicine* and *Statistics and Computing*.

Focusing on journals outside the "A class" list (Fig. 16), it appears that *Journal of Applied Statistics*, *Metron* and *Statistics and Probability Letters* have constantly attracted a good number of papers, confirming their appeal among the Italian statistical community.

On the other hand, the *Electronic Journal of Applied Statistics and Communication in Statistical: Theory and Methods* has suffered a consistent decrease in attraction since 2015.

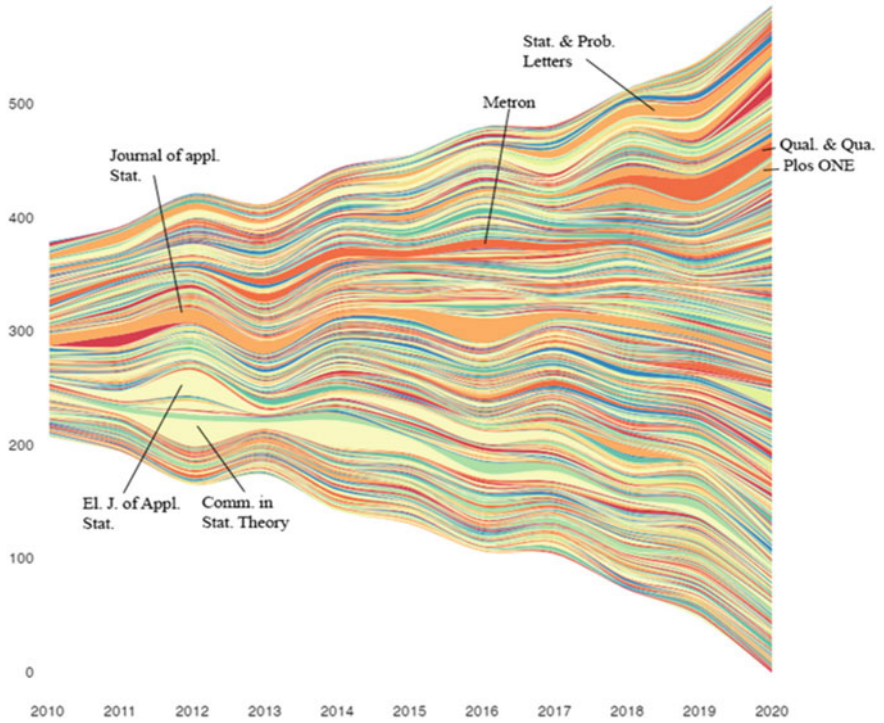


Fig. 16 Streamgraph plot of the number of papers published in journals out of "A class" for the academic recruitment field 13/D1

Similar comments apply to Figs. 17, 18, 19 and 20, containing the streamgraphs related to the rest of the academic recruitment fields.

3 Conclusions

In the paper, we analyzed the research production, in terms of the number of papers published in journals ranked in the Scopus database, during the last ten years by researchers in Statistical Sciences working in the Italian universities. The aim was to investigate the potential effect of the ranking of journals as scientific and "A class," carried out by ANVUR on the habit of publication of the Italian universities' assistant, associate and full professors, classified by academic recruitment fields pertaining to Statistical Sciences. The analysis took into consideration also the consequences of the revisions of the "A class" ranking of journals, as well as an evident effect related to the periods in which the different sessions of the national qualifications were held.

The analysis was carried out mainly with reference to a classification of journals by ASJCs, used in the Scopus database, instead of directly on journals. This choice

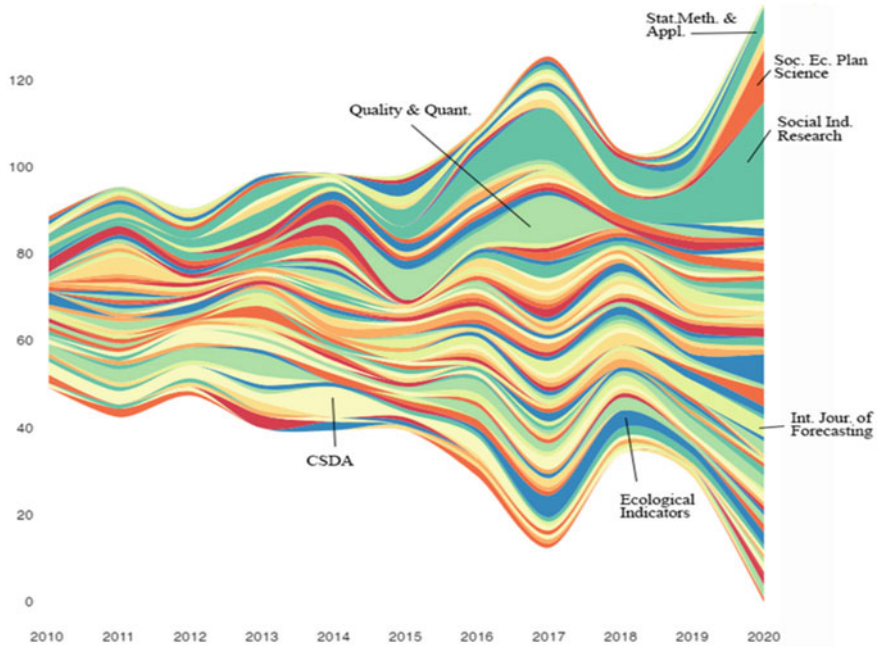


Fig. 17 Streamgraph plot of the number of papers published in "A class" journals for the academic recruitment field 13/D2

is motivated by the fact that this classification is used in the algorithm made up by ANVUR to select the A class journals.

A number of observations arise from this study. The choice of mainly using graphic visualization tools allowed us to provide an image of the evolutions of the publication habit of academic researchers in Statistical Sciences. In particular, both the total number of papers published and the number of papers in "A class" journals have increased over time. However, this trend does not seem to be associated with growing interest in journals traditionally considered appealing by researchers in the field. In the meantime, it seems that among the academic statisticians emerged instead an attitude to orient their publications towards journals that became part of the "A class" list in the recent past. This is particularly evident in the case of the academic recruitment field 13/D1 (Statistics), where we observe an increase over time in the number of papers published in "A class" journals classified in ASJC like Social Sciences (miscellaneous), Economics and Econometrics, Management Science and Operations Research and Geography, Planning and Development. In addition, Social Sciences (miscellaneous) is the ASJC presenting the highest increase of papers published in "A class" journals.

Another important aspect deserving further analysis is undoubtedly the possible influence on the increase in the number of papers published in some periods of the publication of special issues, a policy pursued by some journals (including those of

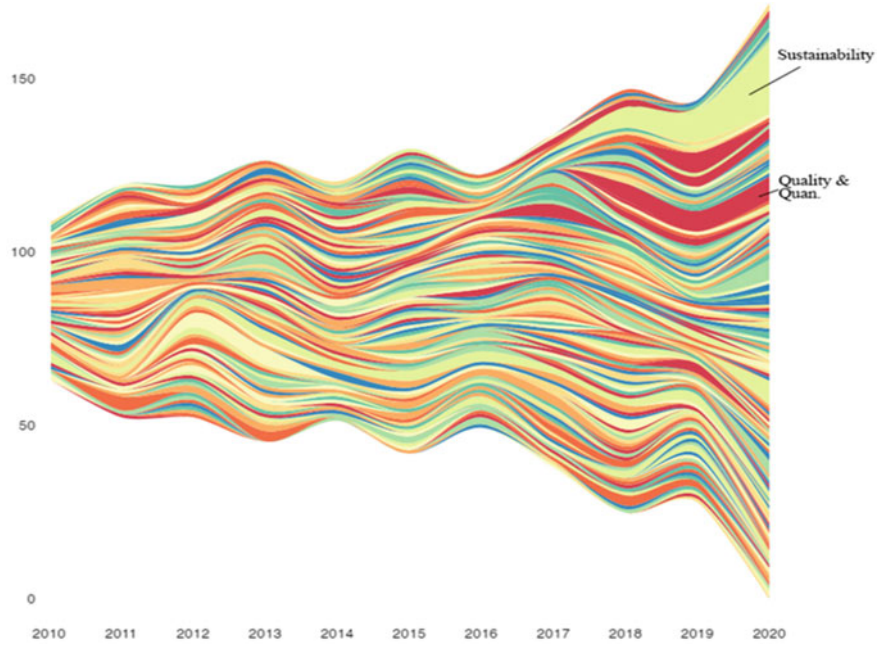


Fig. 18 Streamgraph plot of the number of papers published in journals out of "A class" for the academic recruitment field 13/D2

Social Sciences) to increase their visibility. A future in-depth study of this aspect could contribute to a better understanding of the phenomenon.

Among the open questions, we remind the evaluation of potential effects on the publication habit of different policies adopted by the editorial board of various journals, the number of annual issues published, and the average time from paper submission to publication.

The debate on the influence that the ranking of journals may have on the change of behavior in the publication habit by Italian academics is still open and very active. An international comparison would certainly be interesting in order to understand whether the recent changes observed in Italy are also shared with other countries, where the same ranking of journals does not apply.

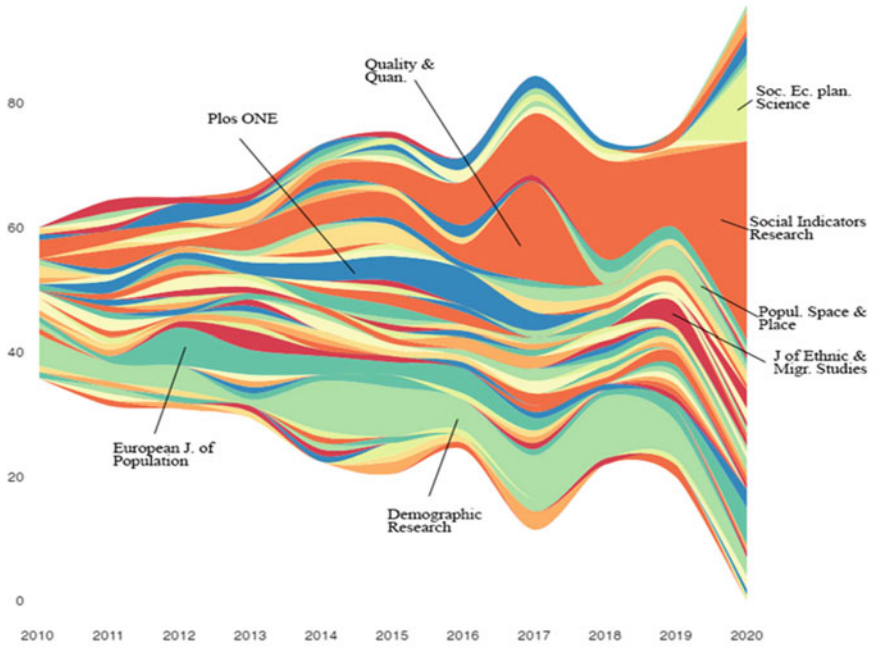


Fig. 19 Streamgraph plot of the number of papers published in "A class" journals for the academic recruitment field 13/D3

Fig. 20 Streamgraph plot of the number of papers published in journals out of "A class" for the academic recruitment field 13/D3

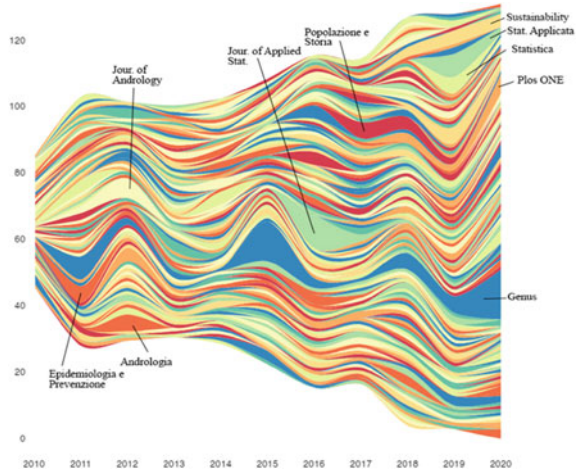


Table 5 Scopus areas

Scopus code	Subject areas	Abbr	Supergroup
1000	Multidisciplinary	MULTI	
1100	Agricultural and Biological Sciences	AGRI	Life Sc
1200	Arts and Humanities	ARTS	Social Sc
1300	Biochemistry, Genetics and Molecular Biology	BIOC	Life Sc
1400	Business, Management, and Accounting	BUSI	Social Sc
1500	Chemical Engineering	CENG	Physical Sc
1600	Chemistry	CHEM	Physical Sc
1700	Computer Science	COMP	Physical Sc
1800	Decision Sciences	DECI	Social Sc
1900	Earth and Planetary Sciences	EART	Physical Sc
2000	Economics, Econometrics and Finance	ECON	Social Sc
2100	Energy	ENER	Physical Sc
2200	Engineering	ENGI	Physical Sc
2300	Environmental Science	ENVI	Physical Sc
2400	Immunology and Microbiology	IMMU	Life Sc
2500	Materials Science	MATE	Physical Sc
2600	Mathematics	MATH	Physical Sc
2700	Medicine	MEDI	Health Sc
2800	Neuroscience	NEUR	Life Sc
2900	Nursing	NURS	Health Sc
3000	Pharmacology, Toxicology, and Pharmaceutics	PHAR	Life Sc
3100	Physics and Astronomy	PHYS	Physical Sc
3200	Psychology	PSYC	Social Sc
3300	Social Sciences	SOCI	Social Sc
3400	Veterinary	VETE	Health Sc
3500	Dentistry	DENT	Health Sc
3600	Health Professions	HEAL	Health Sc

Appendix

Scopus ASJC can be grouped according to the following subjects areas (Table 5).

References

1. ANVUR: Revisione generale delle riviste di classe A per l'area 13. Relazione di accompagnamento, a cura del Gruppo di Lavoro su Riviste e Pubblicazioni Scientifiche per l'Area 13 (Silvia

Fedeli, Marco LiCalzi, Michelangelo Vasta, Enrico Zaninotto) (2016). https://www.anvur.it/wp-content/uploads/2017/09/GdLArea13RelazioneAccompa.pdf
2. Carpita, M.: Valutazione della Ricerca e Classifiche delle Riviste Scientifiche nelle Scienze Statistiche: l'esperienza SIS. Statistica & Società, anno **III**(3), 19–26 (2014)
3. Cocchi, D.: Gli statistici nell'area 13 italiana e nei settori ERC. Statistica & Società, anno **III**(3), 38–40 (2014)
4. Frosini, B.V.: Valutazione della ricerca e valutazione delle riviste scientifiche in ambito statistico. Statistica & Società, anno VI n. speciale, 39–47(2008)
5. Petrucci, A.: Il riconoscimento della scientificità delle riviste: l'esperienza del CUN attraverso la Consultazione Pubblica. Statistica & Società, anno **III**(3), 5–9 (2014)

Trusted Smart Surveys: Architectural and Methodological Challenges Related to New Data Sources

Mauro Bruno, Francesca Inglese, and Giuseppina Ruocco

Abstract In the last few years, Official Statistics have been deeply impacted by the development of smart technologies. This paper summarizes the architectural achievements and the main methodological aspects of the ESSnet (European Statistical System network project) on Smart Surveys, launched at the beginning of 2020. The main goal of the ESSnet is to deliver preparatory work for the development of a European platform, to share und re-use methods and tools for smart data processing. More precisely, the project aims at implementing and testing a common framework for (trusted) smart surveys, through the design of a reference architecture and the development of methodological and technical capabilities within the European Statistical System (ESS). Further, the use of innovative data sources forces National Statistical Institutes (NSIs) to face new challenges, e.g., access to data owned by public and private parties, data processing across multiple NSIs. Privacy preserving technologies are exploited in this paper with the aim to understand their impact on both the architectural framework and technical requirements of the platform.

The expected benefits of developing a shared infrastructure are the decrease of respondent burden, the modernization of statistical processes, as well as the harmonization and enrichment of the statistical output.

Keywords Smart data sources · Trusted smart surveys · Input privacy techniques

M. Bruno (✉) · F. Inglese · G. Ruocco
ISTAT—Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Roma, Italy
e-mail: mbruno@istat.it

F. Inglese
e-mail: fringles@istat.it

G. Ruocco
e-mail: giruocco@istat.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_31

1 Introduction

In the last few years, Official Statistics have been deeply impacted by the development of smart technologies. The use of data captured or generated by smart devices for statistical purposes allows improving the design and the accuracy of traditional social surveys. In order to guarantee the reliability of the statistical output, these new scenarios foster an analysis of the statistical process with respect to the different types of smart data sources.

Smart surveys are based on the combination of several collection modes, such as: (i) data provided by the respondents (active data); (ii) data collected passively by the device sensors (e.g. activity trackers, microphone, GPS, camera, etc.); (iii) data gathered by trusted third parties, (e.g. public authorities, private organizations) and shared by the respondents. Trusted smart surveys (TSS_u) correspond to an enhanced version of the smart survey model, based on methods and tools assuring trustworthiness and facilitating respondents' participation. The key features of TSS_u are: protection of personal data, also through the adoption of privacy-preserving techniques, as well as process auditability and transparency [1].

This paper summarizes the architectural achievements and the main methodological aspects of the ESSnet (European Statistical System network project) on Smart Surveys [2], launched at the beginning of 2020. The main goal of the ESSnet is to deliver preparatory work for the development of a European platform, to share und reuse methods and tools for smart data processing. More precisely, the project aims at implementing and testing a common framework for (Trusted) smart surveys, through the design of a reference architecture and the development of methodological and technical capabilities within the European Statistical System (ESS). The expected benefits of developing a shared infrastructure are the decrease of respondent burden, especially in social surveys, the modernization of statistical processes, as well as the harmonization and enrichment of the statistical output.

2 Methodological Aspects of a Trusted Smart Survey

Smart surveys offer new opportunities for developing social surveys as they aim to collect new data sources through devices (smartphones, tablets, wearables) that use sensors to provide information about themselves or their surroundings. The measurement capabilities of mobile devices can supplement or potentially even replace self-reports in surveys: sensor data collected passively (e.g. location, motion, activity trackers) and respondents' activities on smartphones (e.g. taking pictures, scanning receipts) increase available data sources.

Connecting new data sources—sensor and app data—with self-reports in social surveys represents an added value of smart surveys compared to traditional surveys or digital data. In fact, the integration of different data sources can mitigate the surveys

and digital data weaknesses. Smart surveys form a bridge between primary (survey) data collection and secondary (big) data collection.

However, there are multiple challenges to collecting sensor and app data: participant selectivity, (non) willingness to provide sensor data or perform additional tasks, privacy concerns and ethical issues, quality and usefulness of the data, etc. These aspects have consequences in terms of both representation (selection) and measurement errors. The application of the total survey error framework [3] can provide a useful tool to guide methodological and practical decisions in sensor-app-based data collection. However, it needs to be redefined taking into account the hybrid forms of data collection and the device features for collecting, linking or processing data (device intelligence, internal and external sensors, data donation).

Inevitably, smart surveys lead to new sources of representation and measurement errors, that need the development of new data collection strategies aiming to prevent and control possible sources of error and new methodologies in the assessment and correction of different types of error.

2.1 Representation and Measurement Errors

Representation errors are determined by the availability or not of a smartphone or other mobile devices by the individuals selected in the sample (coverage error), or by their willingness to participate (non-response). Participation is influenced by technological barriers, topic of the survey, duration of data collection, respondent characteristics including privacy and security concerns, and respondent ability with smartphone and its tasks [4]. Consent to participate is required for legal and ethical reasons, but willingness to consent varies per type of sensor and depends on the context and purpose of the measurements. Increased intrusiveness of a sensor measurement can seriously affect the response.

Non-response can occur at many stages, not only from the consent to participate, to download and install an app or device, but also to use the app (whether actively or passively), to capture and transmit data, often repeatedly over a period. The question of non-response becomes more complex as the additional tasks that can be performed increase. Activities can vary in the degree of involvement of the participants, the level of burden, the sensitivity of the data collected, the technical requirements (e.g. battery usage or data transmission volume).

The task of downloading an app for data collection might involve further potential self-selection effects, as it requires additional steps from participants. Mechanisms of respondents' willingness to share sensor data depend on control over data collection, smartphone ability and privacy concerns. Willingness to share may be greater for activities where participants have control over what data are collected and when.

The growing rate of smartphone usage does not solve the coverage problem, if those who use smartphones are different from those who do not in the characteristics of interest, and if the respondents who are willing to engage in specific tasks (e.g. install apps, share sensor data) differ from non-willing smartphone users.

Furthermore, there are differences between those who use smartphones, due to the existence of different operating systems. iPhone owners differ significantly from other smartphone owners in their attitudinal and behavioural characteristics, and these differences cannot be corrected by weighting based on socio-demographic information [5].

Sensor data introduce significant changes to measurement, starting from the definition of the concepts themselves. In fact, measurement errors can be caused by incorrect starting concepts, and/or by the inadequate operational definition of the variables. The conversions of the theoretical concepts do not adequately measure the concept that was to be analysed, or only partially measure it.

Measurement errors in sensor data can occur during the collect phase and in the processing phase. Within the data collection process, measurement errors are generated by different sources, by the respondents' behaviour or by the sensors themselves. Operational errors are determined by the respondents who may incorrectly initialize the measurements or use the devices wrongly. Sensor measurement from smartphone differs by operating system. While iPhones and Android devices usually have the same or very similar embedded sensors, the way these sensors interact with the operating system (OS) (e.g., how often measurements are taken with a sensor), and whether and how external apps are allowed to interact with the sensors, differs by OS. In practice, it is very difficult to develop research apps that work exactly the same across all brands of devices. Similarly, it is difficult to standardize in-browser sensor measurement. Different sensor-equipped devices can produce different results, raising the issues of comparability. The speed of innovation in sensor measurement poses further threats to comparability of measurement over time.

The quality of sensor measurement can be affected by sensor inaccuracy (imprecision, time inequivalence, device inequivalence). Depending on sensor quality and age, sensors may produce systematic and random measurement errors. Systematic errors occur when the sensor measurements deviate from known absolute levels over time (drift). Periodic recalibration is needed to avoid time-dependent systematic errors, but incorrect calibration can produce systematic errors themselves. Instead, random deviations of sensor measurements over time produce noise.

Measurement errors are reporting or data capturing errors and are not only technological, but also human introduced errors. During the processing phase, specification errors may be introduced when sensor data are manipulated, to search for patterns or to explore the accuracy and precision of data, as well as when different sensors are combined. The processing of sensor data is made complicated by the volume of data and the need to adopt processing strategies (such as aggregation or sampling) before their use. In addition, the evaluation and adjustment phase of measurement errors—outliers, noise, missing data—can be time-consuming, especially in the search for appropriate methodological solutions, as in the case of the treatment of missing data.

Sensor data can be missing for short periods of time, due to communication loss or technical issues but, also, for longer periods. The entity of missing data may vary due to smartphone batteries running empty, or a particular sensor, an app, or the device itself when it is turned off by the participant. Measurement challenges can

exasperate the missing data problem, and the collected data will not reflect the true behaviour of an individual. This is the case in which participants install the apps but fail to carry a smartphone everywhere. The strategies of dealing with missing items are very complex because data vary across sensors, depending on the extent and nature of the missingness patterns, and the phenomena under study.

2.2 Smart Data Quality and Management of Error Sources

Representation and measurement errors in smart surveys can be reduced/controlled during the data collection phase, but it is necessary to define at the design phase the best strategies to maximize participation, to prevent measurement errors and to analyse the quality of data.

Data collection strategy concerns the use of contact and reminder strategies, recruitment materials and incentive approaches. The type of interviewer assistance needs to be defined: interviewers can be involved in recruiting interviewees and keeping them motivated; interviewers can be involved from the start or only after a non-response. The recruiting material should be prepared taking into account the activities in which the interviewee is involved. As app data collection requires downloading and registering an unknown app, the recruiting material can include instructions, an overview of basic screens, a landing page, a brief tutorial on how to navigate and possibly a brochure explaining what data is collected and for what purpose (to ensure data confidentiality).

Incentives of different nature (monetary, gamification, feedback) can be used to increase participation or avoid dropping out, but also to counteract the privacy intrusiveness of passive (sensor) data collection. The choice of incentives depends on the burden of the survey for the respondents and on the privacy intrusiveness of sensors. The satisfaction of one type of incentive over another depends on the characteristics and the smartphone skills/habits of the respondents involved in the survey. Providing in-app feedback to respondents that can be instantaneous or postponed to the end of the data collection might motivate more people to participate in the survey, and participants might be more motivated to provide accurate data, so that the feedback is more useful to them. However, personalized feedback might lead participants to change the behaviour that is being measured with the app. In addition, it is costly to implement, and also constrains other design decisions for the data collection.

An important concern of data collection in smart surveys is data quality. While it is true that sensor data acquired passively can lead less measurement errors than self-reports, it is also true that these data are not free from biases. The heterogeneity in sensor quality across smartphone types and the variations in availability of data affect the measurements. Additionally, in case of participatory sensing, the biases that are generated for traditional surveying have to be taken into account. During the data collection phase it is very important to implement quality checks. Soft and hard checks of plausibility of entered data, notifications of missing data implemented in

an interactive and dynamic model that offers insight into the process operation and improved monitoring are needed.

The acquisition of paradata in smart surveys must be designed considering the methods adopted for data collection (active or passive), the functionality of the app developed, the type of device used (smartphone, wearable) and other features performed. The choice of indicators to assess the overall smart survey performance is a complex process, requiring more empirical evidence about the relevance of the information that can be acquired from a device.

Paradata can mitigate survey errors as they are useful: to detect no activity signal or to get information on each contact over time; for tracking, through logs, information on how certain functionalities of the app were used (e.g. how often did the respondents open the insights page); to detect insight in technical difficulties in using the survey app related to the device. The implementation of log files through which an app records and stores events is a complex task, because all the problems that may arise during the collection phase should be taken into account in advance. Information concerning which browser is being used, what version, and on which operating system can be acquired through a browser's user agent string (UA).

Furthermore, paradata can be used to have control over what is measured in the app, to perform comparison of expected results and observation over time (diversity in reports, verification of rule-based category, etc.).

To assess data quality for a smart survey, contextual data on the app usage and on performance of sensors are needed. Users' behaviour with apps may vary from user to user, according to their contextual information in different dimensions such as temporal context, work status in workday or holiday, spatial context, their emotional state, Wifi status, or device related status etc. App usage pattern can be collected from built-in sensors and application programming interfaces. By processing sensor data (consistency validation, metadata enrichment), context information are generated for extracting behaviour patterns or a subject's activity.

The contextual data should assist all likely types of representation and measurement errors that one would like to analyse and/or adjust. For representation and sensor data, it is useful to know if the respondent has access to the sensor, has the ability to use the sensor, if the sensor produces missing data, or if there were problems in data transmission. Here, context information is intended to capture the respondent's behaviour/ability (ability to operate the sensor, to handle the sensor according to instructions), the performance of sensor itself (reliability, deterioration, anomalies) and the problems occurred during reading the sensor data. Advanced analytic techniques to discover information, hidden patterns, and unknown correlations among the contexts are necessary.

In defining a general data collection framework for the TSS_u, several dimensions must be taken into account that can affect participants' concerns and data quality (e.g. criteria for sensor selection related to research objectives and logistics, to the evaluation of sensor characteristics, to participant engagement, to human participant protection). Minimize the risk and burden on participants while maximizing the quantity and quality of data is of primary importance. The set of the sensors used can play an important role in the outcome of a survey, as data quality is intrinsically

constrained by the characteristics of the sensors and the interactions of the participants with those sensors.

Data quality needs to be analysed considering the type of sensor and analytic goals involved, but also the specific features of a smart survey. Indeed, a TSS_u can employ device intelligence and internal sensors as well as other smart features, such as access to external sensors (e.g. activity trackers) and personal and public online data, linkage consent. In the TSS_u design, many aspects must be considered, such as: the trade-off between passive and active data to obtain, for example, a high and balanced response and data quality; the right boundary between respondent burden, respondent engagement and data quality; not least the integration of data from different sources and with different quality levels.

New solutions for the collect phase and new methods for processing data, not yet explored in the traditional surveys, are relevant goals of the ESSnet on Smart Surveys. Machine learning algorithms play an important role in smart data acquisition and in processing or mixing/fusing sensor data. Furthermore, machine learning applications perform better with human feedback. Keeping experts, or respondents themselves, in the loop can improve model accuracy and reduce data errors.

3 Modelling an Architectural Framework for TSS_u

The design of the TSS_u platform has started from the alignment with official statistical standards and existing frameworks, specifically:

- The Generic Statistical Business Process Model (GSBPM), the reference standard describing the main steps of the statistical process and their interdependencies [6].
- The Generic Statistical Information Model (GSIM), fostering the standardization of data and metadata structures, as well as of the other information objects involved in data processing [7].
- The Big Data REference Architecture and Layers (BREAL), an architectural framework resulting from the ESSnet Big Data II, to support the Statistical Institutes in the development of Big Data infrastructures [8]. BREAL is compliant with several architectural standards, such as the ESS Enterprise Architecture Reference Framework (EARF) [9].

Further, the design of a TSS_u platform is a complex task that involves the analysis of technological, methodological and trust aspects. Therefore, a set of guiding principles have been defined to support the definition of the conceptual and logical framework for the TSS_u are listed below:

Architectural principles

- The TSS_u platform should provide software solutions for the management of both traditional surveys (e.g. traditional data collection tools) and new data sources (e.g. sensors, smart data provided by private stakeholders)

- The TSS_u platform should provide configurable application services for particular instances of TSS_u surveys, beyond specific national requirements
- Reuse of available software solutions and development of intuitive interfaces to support end users.

Methodological principles

- The TSS_u platform should provide methods to manage smart missing data and validate input data, also through the interaction with respondents and/or smart devices
- The TSS_u platform should use smart algorithms to reduce response burden and avoid interfering with respondents' activities
- The TSS_u platform should interact with smart devices through algorithms, in order to minimise the consumption of battery power and communication bandwidth
- The TSS_u platform should foster the adoption of customized incentive schemes, to increase the participation rate.

Trust principles

- Hardware and software components developed for the TSS_u platform should guarantee security, confidentiality protection, quality assessment and process auditability
- The TSS_u platform should ensure privacy preservation, through the adoption of input privacy preserving techniques (e.g. Secure Multi-Party Computation, Homomorphic Encryption, Trusted Execution Environment).

In order to share and reuse existing software components, the design of a reference architecture for TSS_u should consider both traditional survey pipelines and the emerging issues arising from new data sources. The next paragraphs provide an overview of the business and the application layers, as well as the operational model, describing different scenarios for the deployment of the software solutions implemented. From the architectural perspective, the analysis of the business layer allows to identify and standardize the core tasks (*What*) to perform for collecting and processing smart data sources. The design of the main process steps is essential to model the application layer that details the software solutions (*How*) to develop for a common infrastructure.

3.1 TSS_u Business Layer Modelling

The design of the business layer allows identifying and standardising the main steps of TSS_u, regardless of the survey domain, or the smart data source. The following analysis, based on the mapping between GSBPM phases and smart data pipelines, summarizes the main steps to acquire and process smart data. More in detail, in the first GSBPM phase "Specify needs", the definition of a smart data strategy must take into account:

- **Type of data provider** (respondents and/or third parties), to set up the technical and legal solutions to manage data confidentiality and privacy issues. According to the "push computation out" approach, the protocols for smart data provided by third parties may involve the use of Input privacy preserving techniques, or the quality assessment of acquired data through a set of quality indicators (more details on Input privacy aspects will be provided in section "Impact of trust on platform requirements")
- **Type of smart data collection**, either passive through the sensors (e.g.: GPS, accelerometer) or active by the respondent
- **Data storage environment**, according to the type of smart data source (e.g.: relational, NoSQL, Json)
- The **monitoring system**, to assess the accuracy of smart data during data collection. The relevance of an early assessment of smart data quality is one of the main issues underlined during the project by Work package 2, conducting pilot surveys to test the feasibility of TSS_u
- **Incentive schemes** for increasing the response rates.

The second GSBPM phase, the "Design phase", allows to plan in detail the main activities to perform and is focused on the following aspects:

- The statistical variables which can be derived from smart data.
- Survey design (mixed mode, adaptive design, data missing planned design) and sampling methodology.
- Contact and recruitment strategies to increase smart survey participation.
- Strategies to prevent and reduce smart survey errors.
- Methods and algorithms to transform smart data in statistical output.
- Metadata management and quality indicators for data and process assessment.
- Data governance, confidentiality and security management.

In the proposed architectural framework, the GSBPM phase explored in depth is the "Collect" phase. Such phase is closely related to the BREAL business functions, belonging to the "Development, Production and deployment" subset:

- Acquisition and Recording, that is the ability to gather Big Data sources
- Data Wrangling, corresponding to the ability to transform the source data format into a particular target format
- Data Representation, as the ability to assign a structure to unstructured or partially structured data
- Modelling and Interpretation, related to the ability to develop and test specific methods and models to process Big Data.

In BREAL, business functions describe behaviours and principles for the management of Big Data sources. The mapping between these business functions and the GSBPM phases aims at highlighting the impact of smart data on the statistical

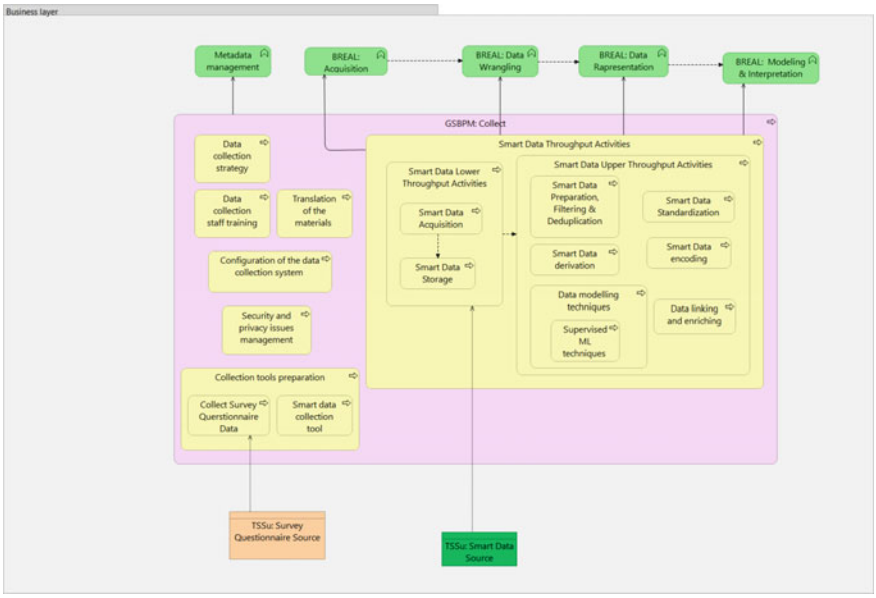


Fig. 1 Smart data "Collect" phase

process. The process steps to perform for smart data gathering, described through the ArchiMate language,[1] are detailed in the figure below (Fig. 1).

The "Collect" phase includes a set of tasks, the Smart Data Throughput Activities, to transform smart data sources in statistical output. Depending on the type of sensor device, data provider and data processing, such activities can be divided in two main subsets: Lower throughput activities, related essentially to smart data acquisition and recording, and Upper throughput activities, grouping the following tasks:

- Data preparation, filtering and deduplication, grouping pre-processing steps to select the relevant information to be processed in the subsequent tasks
- Data standardisation, to convert the source format in a target format, particularly in case of passive data acquisition
- Data derivation, to transform unstructured data sources to structured information. While the previous sub-step concerns data format, data derivation allows to transform data content (e.g., definition of new variables), through a set of rules and/or algorithms
- Data encoding, for the conversion of categorical data in binary or numeric format
- Data modelling techniques, a set of methods, such as machine learning techniques, to derive statistical information from smart data
- Data linking and enriching, to integrate questionnaires data provided by the respondents and the statistical output derived from smart data.

[1] ArchiMate is an open and independent language for architectural modelling, compliant with Enterprise Architecture standard and available from: https://www.archimatetool.com/.

In order to model the TSS_u platform, other relevant tasks to be considered in addition to Smart Data Throughput Activities include:

- Harmonization of a common collection strategy with national peculiarities
- Recruitment and training of the data collection staff
- Set up of collection tools
- Configuration of the data acquisition system
- Management of security and privacy issues during data gathering.

3.2 TSS_u Application Layer Modelling

The analysis of the main steps of the smart data pipeline is the starting point for modelling the application services and components to be implemented for executing the tasks analysed in the business layer. More in detail, the application services needed to execute the aforementioned activities can be grouped in the following subsets:

- **Smart data tools**: application services to standardize the development of collection tools for smart data gathering, according to the type of data provider and the data collection strategy. This subset of services may also provide some functionalities to create incentives to increase the respondents' cooperation
- **Smart data acquisition**: application services grouping a set of functionalities to collect and store smart data actively sent from respondents, as well as data acquired from smart devices or third parties' platforms
- **Smart data processing**: application services to extract statistical information from smart data. Modelling a standardized pipeline would enable the reuse of implemented solutions, although some components must be specialized, according to the peculiarities of some smart data sources
- **Smart data monitoring**: application services to check the fieldwork and assess the quality of smart data gathered, also through the analysis of paradata[2] and contextual data
- **Smart metadata management**: application services to store and manage the process metadata produced and used by the smart data pipeline, to facilitate process auditability and monitoring
- **Input Privacy Preserving Techniques**: usually executed in a distributed computing environment, and applying algorithms and protocols to deal with privacy issues during the data acquisition and processing.

The figure below shows the application services, modelled following a modular approach based on the reuse of the existing software solutions (Fig. 2).

[2] For the definition of paradata, see: https://en.wikipedia.org/wiki/Paradata.

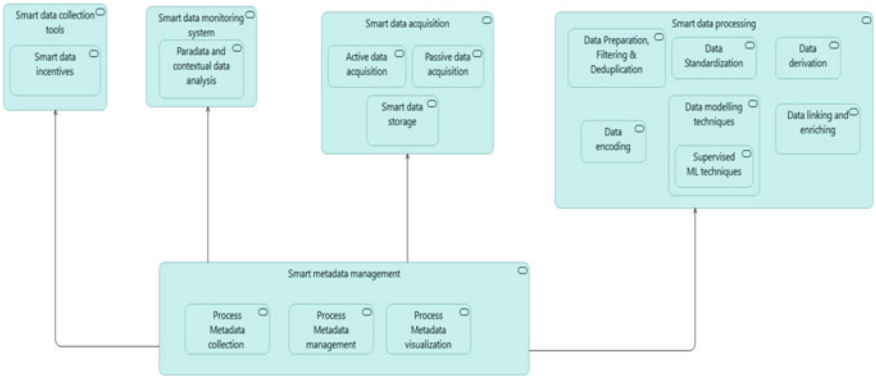


Fig. 2 TSS_u platform application services

3.3 TSS_u Operational Models

The design of the business and architectural layers enables the definition of a strategy for the deployment of the modelled solutions, in relation to some initial assumptions.

The following approach is based on the BREAL operational model, relying on the ESS EARF standard, and classifying the application services in the following subsets:

- **Autonomous services**, implemented in the NSI's local environment, without harmonised implementations across countries
- **Interoperable services**, developed with different back-end and similar service interfaces between the NSIs
- **Replicated services**, if the same application service is provided to different NSIs
- **Shared services**, when available application services are delivered through a platform and accessed by several NSIs.

The infrastructures and platforms dealing with data hosting and management can be classified, with respect to the location and the owner, in the following subsets:

- Local platforms, developed and managed by NSIs
- Local platforms, managed by external data providers (e.g. private cloud) and accessed by NSIs to acquire processed data
- Shared platforms, implemented by a third party and providing shared statistical services for smart data processing (e.g. hybrid cloud).

The following figure shows BREAL operational model that includes also the coverage of data sources, classified in Local, European and Worldwide data [10] (Fig. 3).

In relation to the TSS_u platform, the operational scenarios result from the combination of the following dimensions:

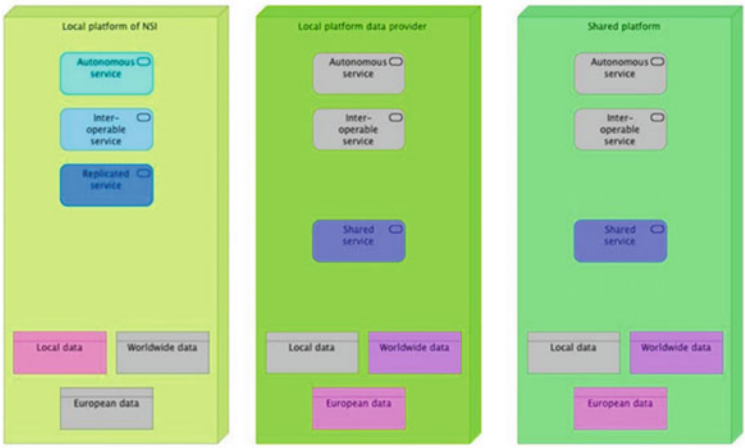


Fig. 3 BREAL operational model

- **Type of data acquisition**. The deployment of application services for smart data collection may vary according to passive or active data acquisition
- **Type of data provider**. The platform could offer software solutions to acquire data from the respondents or third parties through several modes, in compliance with privacy preserving requirements
- **Data Processing and storage environment**. Service deployment must take into account different requirements, according to the different environments where data are stored and processed (In-app/NSI's local environment/TSSu Platform/Third parties). A good practice is to process data at rest, in the storage environment
- **Service deployment** (Autonoumus/Interoperable/Shared/Replicated). Depending on available skills and capabilities, the application services offered by the TSS_u platform could be executed centrally, in the common infrastructure, or locally. If the software solutions must be customized to meet national requirements, the TSS_u platform may provide interoperable services to be configured accordingly.

As an example of an operational scenario, the following figure shows a possible interaction between the NSI's environment and the TSS_u platform in case of the application of Input Privacy Preserving Techniques on data locally stored in NSIs' premises. More details on the impact of such techniques on the platform's requirements and architecture will be provided in the following section).

3.4 Impact of Trust on Platform Requirements

The use of new data sources in the statistical production process forces NSIs to face new challenges, e.g., access to data owned by public and private parties, data processing across multiple NSIs. Data is often sensitive, including details about individuals or organizations that can be used to identify them, draw conclusions about their behaviour and health. Privacy-preserving computation technologies have emerged in recent years to provide protection against such harm while enabling valuable statistical analyses. These new needs lead NSIs to consider input privacy techniques, with the aim to protect the privacy of data acquired by external parties [11, 12].

Data is vulnerable to leakage both by outsiders and insiders in different phases of a typical data processing pipeline:

- **data at rest** (e.g., when data is stored in a server): in the past, when cyber threats were less advanced, most attention to privacy was devoted to data at rest, giving rise to technologies such as symmetric key encryption.
- **data in transit** (e.g., when communicated over the Internet): when unprotected networks such as the Internet became commonplace, attention was focused on protecting data in transit, giving rise to technologies such as Transport Layer Security (TLS).
- **data transformations** (e.g., when used to compute statistics): more recently, the rise of long-lived cyber threats that penetrate servers worldwide gave rise to the need for protecting data during computation. Input privacy techniques are based on data «transformations» that preserve source data privacy. The main input privacy techniques are: Secure Multi-Party Computation (SMC), Homomorphic Encryption (HE), Trusted Execution Environment (TEE), Federated Learning (FL).

A description of the broad context of the Privacy Enhancing Technologies is out of the scope of the paper. In this section, a use case on the application of Input Privacy Techniques (Federated Learning) highlights the impact of input privacy on the TSS_u architecture and requirements.

Federated learning is an input privacy technique to perform a machine learning (ML) training in a private way. It is applied to train predictive ML models on data that cannot be shared. Federated learning technique is based on distributing the algorithm to where the data is, instead of gathering the data where the algorithm is (decentralized/distributed computation).

The following use case simulates a runtime environment with several NSIs, gathering sensors data by accelerometer, with the aim to perform a private ML training. The specific goal of this case study is to perform a classification task, to predict the human activities starting from accelerometer data.[3] The figure below displays the full federated learning process, implemented according to three main steps:

[3] In order to implement the use case, we used on open-source federated learning framework named "Flower" [13].

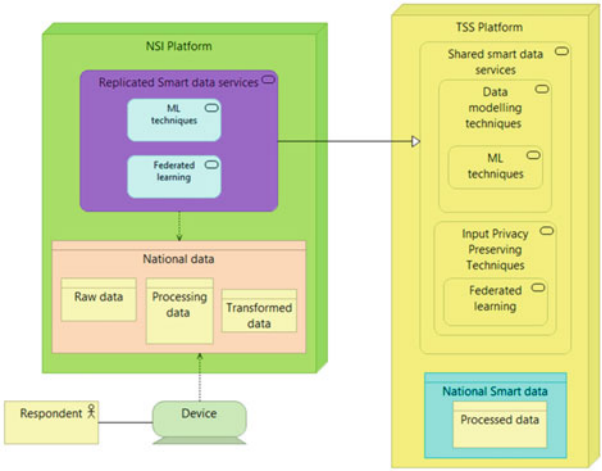


Fig. 4 Input privacy techniques applied through replicated services provided by the TSS_u platform

- **Step 1**: the TSS_u platform provides a service that allows to perform machine learning classification task on accelerometer data ("Shared smart data service" displayed in Fig. 4). The NSIs involved in the process will install the service in their own data center ("Replicated smart data service" displayed in Fig. 4) so that the service will access the data available in the NSI. The TSS_u platform will also provide a FL service, responsible of: (i) the orchestration of the learning process on the server side; (ii) the implementation of the FL strategy (e.g.: average, gradient descent). According to the server inputs, a ML model (e.g., linear regression, neural network, boosting) is chosen to be trained on local nodes and initialized. Then, nodes are activated and wait for the central server to give the calculation tasks.
- **Step 2**: The TSS_u platform orders the selected nodes to execute the training of the model on their local data in a pre-specified fashion (e.g., for some mini-batch updates of gradient descent). At the end of the computation, each node sends its local model to the server for aggregation.
- **Step 3**: The central server aggregates the received models and sends back the model updates to the nodes. It also handles failures for disconnected nodes or lost model updates. The next federated round is started returning to the client selection phase. Step 2 and Step 3 are iterated until a pre-defined termination criterion is met (e.g., a maximum number of iterations is reached, or the model accuracy is greater than a threshold). Then the central server aggregates the updates and finalizes the global model (Fig. 5).

The use case presented provides important feedbacks that should be taken into account in the implementation of the TSS_u platform. Such techniques impact technical and hardware requirements, depending on the specific technique offered by the TSS_u. As shown in the previous example, FL techniques imply the definition of a

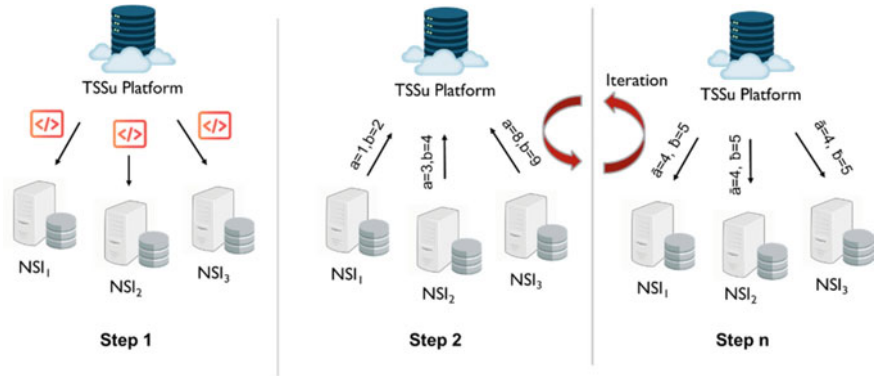


Fig. 5 Steps involved in a federated learning scenario

distributed workflow involving both the central TSS_u platform (that has the role of the server) and the NSI involved in the training process (having the role of nodes). Further, the NSIs should have a runtime environment (e.g. datacenter, private cloud) that allows to install the services offered by the TSS_u platform, and to connect physically to the platform (e.g. network connectivity, communication TCP/IP ports), in order to execute the distributed workflow.

Unfortunately, privacy-preserving computation comes at a cost: current versions of these technologies are computationally costly, rely on specialized computer hardware, are difficult to program and configure directly, or some combination of the above. Thus, both at European level and at National level, scientists may need guidance in implementing such technologies that ensure privacy benefits. The crucial aspect in deploying Privacy enhancing techniques (PETs) is that they have to be deployed as close to the data owner as possible. The best privacy guarantees require that PETs are applied by the data owner, on premises, before releasing confidential data to third parties.

4 Conclusions

The ESSnet has produced relevant results, providing an overview of several issues that have an impact on the technical requirements of the TSS_u platform. Additional exploration of methodological and architectural aspects is required, to gain evidence based insights for improving the achieved outcome. Several Proofs of Concepts (PoC), planned in the second part of the ESSnet project, will address the open issues related to active and passive data gathering, machine learning algorithms for sensor data processing, incentives to increase respondents' participation, data processing environment, metadata management, privacy techniques and technical requirements. The feedbacks resulting from the PoCs and the pilot surveys carried out by Work

package 2 will contribute to the design of an enhanced framework for the TSS_u platform, providing a starting point for future developments.

References

1. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M.: Trusted smart statistics: motivations and principles. Stat. J. IAOS **35** (2019). https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf
2. ESSnet on Smart Surveys (2020–2021). https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys_en
3. Biemer, P.P., de Leeuw, E., Eckman S., Edwards B., Kreuter T., Lyberg L.E., Tucker N.C., West, B.T. (eds.) Total Survey Error in Practice. John Wiley & Sons, Inc., Hoboken, New Jersey (2017)
4. Keusch, F., Struminskaya, B., Antoun, C., Couper, M.P., Kreuter, F.: Willingness to participate in passive mobile data collection. Pub. Opin. Quart. **83**, 210–235 (2019)
5. Struminskaya, B., Lugtig, P., Keusch, F., Höhne, J.K.: Augmenting surveys with data from sensors and apps: opportunities and challenges. Soc. Sci. Comput. Rev. **1–13** (2020). https://doi.org/10.1177/0894439320979951
6. Generic Statistical Business Process Model (GSBPM) v. 5.1. January (2019). Available from: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1
7. Generic Statistical Information Model (GSIM) v. 1.2 March (2021). Availabel from: https://statswiki.unece.org/display/gsim/GSIM+v1.2+documents
8. ESSnet on Big Data II, Work Package F, Deliverable F1. (2018–2021). https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf
9. ESS Enterprise Architecture Reference Framework (EARF), September (2015). Available from: https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en
10. Scannapieco, M., Bogdanovits, F., Gallois, F., Fischer, K.G., Paulussen, R., Quaresma, S., et al.: BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer (2021). Version 2021-03-31. Edited by EUROSTAT
11. Ricciato, F., Giannakouris, K, Wirthmann, A., Hahn, M.: Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. SIS (2020). https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/Pearson-SIS-2020-atti-convegno.pdf
12. Ricciato, F., Bujnowska, A., Wirthmann, A., Hahn, M., Barredo-Capelot, E.: A reflection on privacy and data confidentiality in official statistics. In: ISI World Statistics Congress (2019). https://www.bis.org/ifc/events/isi_wsc_62/ips177_paper3.pdf
13. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., Lane, N.D.: Flower: a friendly federated learning research framework (2020). arXiv preprint arXiv:2007.14390

Web Surveys: Profiles of Respondents to the Italian Population Census

Elena Grimaccia, Alessia Naccarato, and Gerardo Gallo

Abstract Identifying the profile of the "web respondent" can help survey designers to promote the participation in web-based surveys, with the aim of enhancing timeliness and reducing costs of data collection. This study reports results from a mode comparison, between a Computer Assisted Web Interview and a Computer Assisted Personal Interview. The aims are to assess the familial and geographical characteristics corresponding to a greater probability of choosing to respond to a web survey, and to identify the web respondents' profile. Logit models with different specifications have been estimated on the probability of answering via web, based on the 2019 Italian population census data. Regional fixed effects, geographical covariates referring to the municipalities, and interactions were all included in the model to control for the variability of the territories. Results show that the households with a lower level of education, composed by foreigner members, residing in the South of Italy, or in small municipalities, present a lower probability of answering to a web survey, and can be made subject of specific actions in order to increase the share of web respondents.

Keywords Web surveys · Respondent profiling · Logit model · Italian permanent population census

1 Introduction

For statistical surveys, web-based interviews have brought considerable advantages, such as the reduction of costs, the containment of the interviewer effect [3, 25, 26],

E. Grimaccia (✉) · G. Gallo
ISTAT - Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Rome, Italy
e-mail: elgrimac@istat.it

G. Gallo
e-mail: gegallo@istat.it

A. Naccarato
Department of Economics, Roma Tre University, Rome, Italy
e-mail: alessia.naccarato@uniroma3.it

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_32

and a lower participation burden on respondents [11]. These advantages are even more valuable for official statistics because of large sample sizes, very large number of variables to be collected, as well as the high quality standards they must ensure. These issues assume particular relevance in the case of the population census, which represents one of the most important surveys carried out by National Statistical Institutes (NSIs). Moreover, the emergency due to the coronavirus pandemic in 2020 and 2021, and the impossibility of carrying out field operations have further boosted the necessity for web-based interviews.

Even if the use of the internet is quite widespread in Europe [18], the respondents' attitudes toward web interviews is still not granted. In this framework, the knowledge of the interviewees' profile is useful to define strategies to enhance the participation of respondents to interviews via web [4, 7]. During the survey process, decisions have to be made both to contact the eligible respondents and to solicit the compilation of the web questionnaire [5]. The interaction of the characteristics of the respondent with the decisions of the survey designer influences the response rate and the success of the survey [5, 15].

In Italy, the Permanent Population and Housing Census (PPHC), started in 2018 by the Italian National Institute of Statistics (Istat), currently uses a mixed mode data collection in which respondents may choose to fill in the questionnaire via a Computer Assisted Web Interview (CAWI) or in the traditional way. The availability of such data allows to study the different profiles of CAWI and not CAWI respondents.

In this study, then, we analyse the determinants that influence the cooperation of the respondents to the Italian population census in a CAWI mode survey, in order to point out the specific characteristics of the population, which could enhance the effectiveness of the actions of survey designers to improve CAWI participation [6].

Identifying the profile of the CAWI respondent households (HHs) and the characteristics of the HHs for which a CAWI response is most probable is useful also in order to design successive editions of the survey, since the profiling is based on structural economic and social characteristics of the population that do not change quickly over time.

The profiles of the CAWI respondents are also analysed according to their territorial distribution, since geographical differences are particularly significant, being related with a number of factors of economic and social development [2]. Data collection strategies, that have to adapt to the geographical imbalances, will therefore benefit from the analysis proposed in the paper.

2 The Italian Permanent Population and Housing Census

The population census provides official data for the usual resident population living within the national borders [30].

In recent decades, development in information technology and the ever-increasing availability of administrative data have led several European countries to develop innovative methods for the population census, and data are collected using registers

and other administrative sources, together with information from sample surveys [9]. Among the 32 countries of the European Union and European Free Trade Association, a large majority has planned a register-based census for the 2020 round, while only ten will continue with a traditional census [31].

The combined census approach takes full advantage of the use of administrative sources and sample surveys: cost reductions, gain in quality and a higher probability of reach elusive populations like the homeless and irregular or unregistered migrants. Moreover, administrative data are useful in assessing the quality of surveys, adjusting coverage errors, providing covariates, completing information, and also replacing surveys [10, 24].

In Italy, the 2011 population and housing census concluded a long period in the history of official statistics, that has foreseen census of the population every ten years, since 1861 [20]. Starting from 2018, the Permanent Population and Housing Census, an annual sample survey, has been implemented, marking the definitive transition from the traditional "door-to-door" enumeration to a "register-based" system [13]. The PPHC is a complex statistical process, exploiting and integrating the information derived from registers with those collected in surveys on socio-economic variables (Fig. 1).

The main difference between PPHC and the traditional census is that it does not involve all individuals but statistically representative samples of the population. The information on individuals based on sample surveys is integrated with that from administrative sources so as to ensure greater exhaustiveness and an increase in the quantity and quality of information [14, 22].

Further advantages of these new census methodologies are the greater containment of the census participation burden on respondents, a reduction in the overall costs of the survey, and better timeliness [4, 31].

It is worth mentioning that, as in traditional censuses, participation in the PPHC is compulsory by law and the violation of this obligation is subject to sanctions.

The design of the PPHC is based on two different yearly sample surveys: an Areal survey and a List survey. The Areal sample survey results are used to update

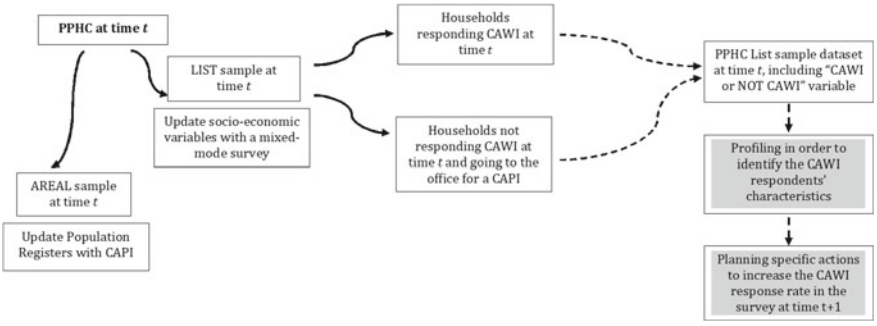


Fig. 1 PPHC data collection design

Table 1 Municipalities and HHs by population size of municipalities—Italian Permanent Population and Housing Census, 2019

Demographic size of Italian municipalities (inhabitants)	Sample group for Areal survey		Sample group for list survey	
	Municipalities	HHs	Municipalities	HHs
Up to 5,000	1,514	137,812	1,028	190,458
From 5,001 to 20,000	816	106,815	815	288,360
From 20,001 to 50,000	376	52,190	376	184,593
From 50,001 to 100,000	98	97,396	98	120,149
From 100,001 to 250,000	34	35,927	34	94,226
More than 250,000	12	22,561	12	73,253
Total	2,850	452,701	2,363	951,039

Source ISTAT (2021)

the Population Registers, while the List survey is carried out in order to collect socio-economic variables [22].

The Areal survey is a "door-to-door" enumeration, which is conducted directly to a specific address based on a specific archive [23]. Instead, with regard to the List survey, HHs receive a letter inviting them to fill in the online questionnaire using the CAWI technique. The HHs in the List sample that are not available to fill in the online questionnaire can opt for a "face-to-face" CAPI interview with the support of an interviewer. The availability of the information on these different survey modes—accounted for in the List sample—allows us to distinguish the HHs which fill in the questionnaire via the web (CAWI respondents) from those who prefer the traditional interview. Thanks to this information, it is possible to estimate the structural characteristics that are associated to the preference to respond online, in order to define the profiles of each type of HH, and to study their geographical distribution. Therefore, for the purposes of our analysis, from now on we will focus on the List survey that represents about 68% of the total PPHC sample. All in all, the Areal and List surveys are based on a yearly sample size of about 450,000 HHs and 950,000 HHs, respectively (Table 1). Therefore, the sample is composed every year of about 1.4 million HHs. Around 1,100 municipalities are involved every year in the surveys, while approximately 1,700 participate once every 4 years. This procedure ensures that—in a period of four years—all municipalities (7,904) have participated at least once in the PPHC.

3 Data

The overall CAWI response rate for the 2019 Italian PPHC is 49.9%. The share of HHs that answer to a survey conducted with a CAWI methodology varies according to the residence of the HH: in the Northern area of Italy the share of CAWI's respondents

is close to 59%, while in the "Mezzogiorno" area (Southern area of Italy plus the two major Islands) the share of CAWI respondents is stuck at 36% (Table 2).

The CAWI respondents—who are called to fill the questionnaire themselves, with huge costs' saving for NIS—did not actually need any help during the survey for the 80.4% of interviews. The 16.7% had the support of relatives or friends, while only the 1.4% called the help line, and only the 1.2% of respondents needed to ask for help to other agencies or institutions.

Conceptual frameworks of survey participation have identified a number of key factors that influence web response [4, 15, 21, 27, 28]. The PPHC List sample provides a unique opportunity to study these features in more detail, providing information on HHs and individuals within HHs. In this study, the features analysed are the HH size, the HH citizenship, the youngest HH member age, and the HH highest educational level. The 'HH size' is the number of people in the same dwelling. The 'HH citizenship' indicates the citizenship of its members: all foreigners, all Italians, or a HH including different citizenships among the HH members. The 'youngest HH member age' variable specifies the age class of the youngest member of the HH, following the idea that young people in a HH could help older members in the use of technology. The minimum age admitted is 18 years old, since the census questionnaire cannot be filled by underage members. The 'highest educational level' is the highest qualification among the whole family, and it is classified in primary, secondary and tertiary, grouping the (0, 1, 2), (3, 4), (5, 6, 7, 8) levels of the International Standard Classification of Education (ISCED), respectively.

In order to take into account the different characteristics of territories that are usually related to the use of Internet [2, 8, 12], also geographical variables are included in the analysis: local capital or metropolitan city, municipality demographic size, municipality urbanisation degree, and altimetric area.

A municipality is a local capital or metropolitan city when it is seat for the local government of the "province" or metropolitan area. This is an administrative division at the European NUTS 3 level, at subregional level [17].

The variable 'municipality demographic size' indicates the number of inhabitants: from the lowest class corresponding to "up to 5,000 inhabitants", to the highest class, corresponding to municipalities with "more than 250,000 inhabitants".

Table 2 Distribution of CAWI and not CAWI responses by geographical area—2019

Geographical area	Cawi	Not Cawi	Total
North	58.91	41.09	100.00
Center	53.77	46.23	100.00
Mezzogiorno	36.55	63.45	100.00
Italy	49.99	50.01	100.00
Pearson chi2 = 3.4e + 04 Pr = 0.000			
Kendall's tau-b = −0.1849 ASE = 0.001			

Source ISTAT (2021)

Municipalities are classified into three degrees of urbanisation: cities or large urban areas, corresponding to densely populated areas (defined as clusters of cells of 1 square km contiguous, with a density of not less than 1,500 inhabitants per square km and a population of not less than 50,000 inhabitants); small urban areas, with an intermediate density level (clusters of contiguous cells with a density of not less than 300 inhabitants per square km and a population of not less than 5,000 inhabitants); sparsely populated areas or rural areas, defined as single cells (rural) not classified in the previous groups [19].

The altimetric area indicates the altitude above sea level, classifying municipalities in mountain, hill, and plain areas. The mountain and hill areas have been divided into inland mountain and inland hill areas and coastal mountain and coastal hill areas, in order to take account of the moderating action of the sea on climate. The altimetric zone variable was considered since it could affect the choice to answer online due to the different internet access capacity of the mountain areas.

Municipalities belong to regions, the official level of aggregation of Italian municipalities. This aggregation is particularly important to make evident the economic and social differences in Italy, since many policies are developed by regions. Regions can also be aggregated in macro regions: the north, central, and Mezzogiorno areas of Italy. This classification provides area that are similar according to social and economic level of development.

4 Methods

As in Biffignandi and Pratesi [5], and more recently in Maslovskaya et al. [27] and Rivero et al. [29], a logit model has been estimated in order to identify the determinants that are significantly associated to the HHs with the highest probability of responding via the web. To compute the probability that a family responds in CAWI mode, a logit model, in which the endogenous variable is the dichotomous variable that assumes the value 1 if the family responded CAWI and zero otherwise has been employed [1, 5].

The dependent variable used in the models is, then, the survey mode chosen by the respondent, a binary response Y_i defined as follows:

$$Y_i = \begin{cases} 0 & Not\ CAWI\ response \\ 1 & CAWI\ response \end{cases} \tag{1}$$

The exogenous variables are the social and demographic characteristics of the HHs (namely: HH size, HH citizenship, younger HH member age, HH higher educational level), illustrated in Sect. 3.

A binary logistic regression with binary response Y_i as in (1), in which the probabilities $g_i = \Pr(Y_i = 1|x_i)$ are related to a linear predictor $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots$ through the logit function, is the following [33]

$$logit(g_i) = \log\left(\frac{g_i}{1-g_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots \quad (2)$$

Among the exogenous variables of the model is also considered the variable indicating the region (corresponding to the Eurostat classification NUTS2) including the municipality where the HH lives. Taking into account the regions in the model allows to consider the effect of the economic and social differentiations that historically distinguish different areas of Italy. Therefore, in the model regional fixed effects were also considered, as this allows to study the family characteristics that define the CAWI profile, net of the effects due to the region of residence.

The model has also been estimated for the five areas separately in order to verify if any differences exist in the main drivers (Table 3).

The estimated coefficients are comparable since the structure of the survey, the definition of the variables and their categories are the same in each and every territorial area, as in Espinosa and Hennig [16]. The sample is, indeed, representative for each region and of course for macro-regions. This allows to compare the results in every subsample.

In order to plan interventions for the promotion of web surveys on the territory, it is useful to study what are the characteristics of the areas of the country in which it is more important to intervene. This aim has been achieved including in the model also the characteristics of the municipalities, such as the number of residents, the degree of urbanization, and the altimetry (Table 4).

5 Results

Table 3 reports the results (coefficients and odds ratios) of the logit model that identifies the features of the HHs answering via web data collection, and ultimately identifying their profile.

The empirical evidence suggests that among HHs with a higher level of education the odds ratios (OR) of answering via web are significantly higher than for those who have a lower level degree: HHs with a member with a secondary degree (high school diploma) present odds of answering via web that are double compared with those HHs that are composed only by less educated members. Furthermore, HHs with a member with an education of tertiary level (University degree) present odds that are almost four times those of "less educated" HHs. A higher level of education is, then, a strong determinant of the availability to respond via web in every area of the Country.

HHs composed by all foreigner members present a lower probability of using the CAWI option than those with at least an Italian member. In particular, HHs with all Italians members show four times the odds of answering to a CAWI, compared to HHs with all foreigner members. The same result is obtained in all the Italian macro-regions: foreigner HHs should be made target of specific actions aiming at helping them responding via web in every area of Italy.

Table 3 Logit model estimated parameters (coefficients and odds ratios between brackets)

Variable	Italy	North	Centre	Mezzogiorno
HH highest educational level (base = Primary level education)				
Secondary	0.7364 (2.088)***	0.6870 (1.988)***	0.6700 (1.954)***	0.8444 (2.327)***
Tertiary	1.355 (3.878)***	1.3055 (3.690)***	1.2388 (3.451)***	1.4879 (4.428)***
HH citizenship (base = all foreigners)				
All Italians	1.3829 (3.986)***	1.5434 (4.680)***	1.1888 (3.282)***	1.1798 (3.254)***
Mixed citizenship	0.7142 (2.042)***	0.7244 (2.063)***	0.6422 (1.901)***	0.7107 (2.035)***
HH size (base = 1 component)				
2	0.0108 (1.010)	−0.0027 (0.997)	0.0160 (1.016)	0.0385 (1.039)***
3	0.0171 (1.017)*	0.0255 (1.026)*	−0.0001 (1.000)	0.0367 (1.037)**
4	0.1096 (1.116)***	0.1695 (1.185)***	0.0876 (1.091)***	0.1052 (1.111)***
5 or more	−0.2368 (0.789)***	−0.2171 (0.805)***	−0.2316 (0.793)***	−0.2130 (0.808)***
Youngest household member age (base = 18–34 years old)				
35–64	0.0515 (1.052)***	−0.021 (0.009)*	0.0261 (1.027)	0.1342 (1.144)***
65+	−0.0436 (0.957)***	−0.249 (0.010)***	−0.0783 (0.925)***	0.2640 (1.302)***
Trentino Alto Adige	0.0177 (1.024)	0.0240 (1.018)		
Lombardy	0.2814 (1.325)***	0.2776 (1.320)***		
Piedmont	0.0620 (1.063)***	0.0607 (1.063)***		
Friuli Venezia Giulia	0.0379 (1.038)*	0.0416 (1.042)**		
Veneto	0.1160 (1.122)***	0.1101 (1.116)***		
Liguria	−0.0138 (0.986)	−0.0081 (0.992)		
Valle d'Aosta	−0.1673 (0.846)***	−0.1724 (0.842)***		
Tuscany	−0.0602 (0.941)***		*(base)*	

(continued)

Table 3 (continued)

Variable	Italy	North	Centre	Mezzogiorno
Marche	−0.2244 (0.799)***		−0.1592 (0.853)***	
Lazio	−0.2211 (0.801)***		−0.1524 (0.858)***	
Umbria	−0.2091 (0.811)***		−0.1443 (0.866)***	
Abruzzo	−0.5270 (0.590)***			*(base)*
Campania	−1.0489 (0.359)***			−0.4997 (0.607)***
Sardinia	−0.4825 (0.617)***			0.0641 (1.066)***
Molise	−0.9245 (0.396)***			−0.3918 (0.676)***
Puglia	−0.6598 (0.516)***			−0.1144 (0.892)***
Basilicata	−0.9507 (0.386)***			−0.4113 (0.663)***
Sicily	−1.0459 (0.351)***			−0.5005 (0.606)***
Calabria	−1.3369 (0.262)***			−0.7959 (0.451)***
Constant	−1.6673 (0.188)***	−1.6947 (0.184)***	−1.466 (0.231)***	−2.210 (0.110)***

Legend * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$
Source Istat (2021)

Moreover, in HHs where the age of the youngest members is between 35 and 64 years old the CAWI response odds are significantly higher than in HHs where the age of the youngest member is between 18 and 24 years old. The HHs composed by all elderly people present the lowest probability of CAWI response. However, the Mezzogiorno geographical area is the only one where older people accept to answer via web more than HH with younger members.

Finally, the analysis of the size of the HH in terms of number of components offers less univocal results. The probability of CAWI answering is significantly lower in HHs with 5 or more members; the same probability, instead, is significantly higher for HHs with four components, in all the areas of Italy. However, while in the Northern part of the peninsula other HH dimensions do not make such a difference, in the regions of Mezzogiorno also two and three members HHs show a higher probability of CAWI answering.

It is worth mentioning that HHs with at least a member with higher education have a higher probability of accepting the CAWI mode regardless of age or citizenship.

Fig. 2 Regional odds ratios of CAWI response

The profile of the HH with a higher probability of answering via web is that of a family composed by at least one member with a University-level education degree, and of Italian citizenship. This profile is common to every geographical area taken into account.

The ORs referring to the Italian regions are presented in Fig. 2.

Lombardy, Veneto and Piedmont record the higher probability of a CAWI response (Fig. 2). The other Northern regions are found with a slightly lower odds of a response via web. The Valle D'Aosta region presents the lowest probability of answering via web in the Northern area. The Mezzogiorno's regions—except Abruzzo, Puglia, and Sardinia which are aligned with the Central geographical area—present instead the lowest probability of a CAWI survey mode.

Table 4 reports the results (coefficients and odds ratios between brackets) of the logit model that identifies the profile of the HHs answering via web, including the explanatory variables referring to municipalities as well. Considering the characteristics of municipalities, results show that the larger the demographic dimension of municipality, the higher the odds of a CAWI response. Likewise, in highly populated areas such as cities or large urban areas, the odds of a CAWI response are significantly higher than in small towns and rural areas.

Including in the model the variables referring to the characteristics of municipalities provides additional information on the phenomenon, and allows as well for a control of the robustness of the previous model. Considering the estimations of this second model, indeed, the level of education and the citizenship are still the variables that explain the probability of answering via web the most.

Considering that the OR referring to the municipalities with more than 250 thousand inhabitants is equally among the highest, therefore the profile of the web-responding HH corresponds to a family with at least one highly educated Italian member residing in a large city.

Estimating the model in different geographical areas, some specific features emerge. In the Mezzogiorno, HHs living in a local capital present a significant higher probability of answering via web, while in the North and Central areas of Italy, the ORs are higher elsewhere. This result highlights a very different socio-economic

Table 4 Logit model estimated parameters (coefficients and odds ratios between brackets)

Variable	Italy	North	Centre	Mezzogiorno
Household highest educational level (base = Primary level education)				
Secondary	0.711 (2.036)***	1.956 (0.67101)***	1.889 (0.63588)***	2.232 (0.80315)***
Tertiary	1.2876 (3.624)***	1.256 (3.511)***	1.1529 (3.167)***	1.4036 (4.070)***
Household citizenship (base = all foreigners)				
All Italians	4.227 (1.4414)***	1.5845 (4.877)***	1.2507 (3.493)***	1.2371 (3.445)***
Mixed citizenship	0.7487 (2.114)***	0.7444 (2.105)***	0.6706 (1.955)***	0.7632 (2.145)***
Household size (base = 1 component)				
2	0.0177 (1.018)**	0.0055 (1.005)	0.02397 (1.024)	0.0326 (1.033)**
3	0.0361 (1.037)***	0.0455 (1.046)***	0.02227 (1.022)	0.0372 (1.038)**
4	0.13744 (1.147)***	0.1965 (1.217)***	0.1202 (1.128)***	0.1185 (1.126)***
5 or more	−0.2102 (0.810)***	−0.1851 (0.831)***	−0.2003 (0.818)***	−0.209 (0.811)***
Youngest household member age (base = 18–34 years old)				
35–64	0.04737 (1.048)***	−0.02088 (0.979)*	0.0138 (1.014)	0.1203 (1.128)***
65+	−0.0638 (0.938)***	−0.2632 (0.769)***	−0.1068 (0.899)***	0.2364 (1.267)***
Demographic size (base = less than 5,000 inhabitants)				
From 5,001 to 20,000 inhab.	0.0889 (1.093)***	0.0649 (1.067)***	0.1962 (1.217)***	0.1085 (1.114)***
From 20,001 to 50,000 inhab.	0.10775 (1.114)***	0.0643 (1.66)***	0.1778 (1.195)***	0.2249 (1.252)***
From 50,001 to 100,000 inhab.	0.2437 (1.276)***	0.0984 (1.103)***	0.4409 (1.554)***	0.3521 (1.422)***
From 100,001 to 250,000 inhab.	0.2795 (1.322)***	0.2242 (1.251)***	0.6070 (1.835)***	0.2953 (1.343)***
Over 250,000 inhab.	0.5089 (1.663)***	0.3371 (1.401)***	1.0222 (2.779)***	0.4817 (1.619)***
Altimetric area (base = internal mountain)				
Coastal mountain	−0.0814 (0.922)***	0.1072 (1.113)**	−0.2536 (0.776)***	−0.0062 (0.994)
Internal hill	0.03715 (1.038)***	0.0188 (1.019)	0.1024 (1.108)***	0.0168 1.017

(continued)

Table 4 (continued)

Variable	Italy	North	Centre	Mezzogiorno
Coastal hill	0.0255(1.026)*	−0.0701 (0.932)**	0.1063 (1.112)***	0.0389 (1.040)
Flat land	0.0401 (1.041)***	0.0086 (1.009)	0.1287 (1.137)***	0.0306 (1.031)
Urbanization degree (base = cities or large urban areas)				
Small towns or intermediate population density zones	−0.0606 (0.941)***	−0.0839 (0.919)***	0.1061 (1.112)***	−0.0966 (0.908)***
Rural areas or sparsely populated areas	−0.1817 (0.834)***	−0.224 (0.799)***	0.0679 (1.070)	−0.1982 (0.820)***
Local capital or metropolitan city provincal capital/metropolitan city (base = no)				
Yes	0.0274 (1.028)**	−0.0772 (0.926)***	−0.0318 (0.969)	0.195 (1.215)***
Region (base = Emilia Romagna)				
Trentino Alto Adige	0.1855 (1.204)***	0.1162 (1.123)***		
Lombardy	0.3482(1.416)***	0.3103 (1.364)***		
Piedmont	0.1437 (1.154)***	0.1275 (1.136)***		
Friuli Venezia Giulia	0.0900(1.094)***	0.1022 (1.107)***		
Veneto	0.1767 (1.193)***	0.1486 (1.160)***		
Liguria	−0.0277 (0.973)	0.0016 (1.002)		
Valle d'Aosta	0.0445 (1.045)	−0.0128 (0.987)		
Tuscany	−0.0335 (0.967)**		(base)	
Marche	−0.112 (0.894)***		−0.0727 (0.930)***	
Lazio	−0.2448 (0.783)***		−0.2611 (0.770)***	
Umbria	−0.1763 (0.838)***		−0.1462 (0.864)***	
Abruzzo	−0.4404 (0.644)***			(base)
Campania	−1.033 (0.356)***			−0.5584 (0.572)***
Sardinia	−0.3735 (0.688)***			0.1099 (1.116)***

(continued)

Table 4 (continued)

Variable	Italy	North	Centre	Mezzogiorno
Molise	−0.7544 (0.470)***			−0.3151 (0.730)***
Puglia	−0.6818 (0.506)***			−0.2274 (0.796)***
Basilicata	−0.8225 (0.439)***			−0.3554 (0.701)***
Sicily	−1.057 (0.348)***			−0.6085 (0.544)***
Calabria	−1.230 (0.292)***			−0.7731 (0.461)***
Constant	−1.851 (0.157)***	−1.743 (0.175)***	−1.937 (0.144)***	−2.347 (0.096)***

Source ISTAT (2021)

conditions in the rural areas of the Southern Italy, compared with the Northern area, where large cities have—apparently—areas of disadvantage.

6 Conclusions

The advantages of web surveys are widely recognised by national statistical institutes around the world [32]. In countries where online data collection is not fully operational and is not used in all surveys, as is the case in Italy, one of the main difficulties to be faced—in surveys aimed at individuals or HHs—is the reluctance of the population to filling in online questionnaires.

Identifying the characteristics of the respondents, that lead to this resistance to answer via web, and the peculiarities of the municipalities in which the problem is more concentrated would make it possible to intervene on the population to promote web survey and besides to provide technical aid to those who are not able to answer online.

Web surveys still face—at least in Italy—widespread resistance from the population. In order to foster a positive attitude towards web interviewing, an awareness campaign is necessary, and this would be more effective if it targeted a specific population. Hence, there is a need to identify the HH profiles from which a web response is most unlikely. A further consideration should be made with regard to families living in southern Italy, since in this case the geographical factor plays a very important role. Indeed, in the Mezzogiorno geographical area, only 36% of the HHs choose to answer via the web.

This contribution has shown how some structural characteristics of HHs allow us to classify them on the basis of their preference to fill in the population census questionnaire online. The HHs with a higher level of education, composed of Italian

citizens, living in the North and partly in Central Italy are those for whom the probability of answering via the web is higher. Since these characteristics are usually those referring to HHs that historically live in better economic and social conditions, the results indicate that the attitude toward web data collection is favoured by those characteristics that other studies identify as less fragile conditions.

The participation to web survey is more common in large cities and in high population density urban areas. However, some fragility appears in local capital in the North of Italy.

This study presents detailed characteristics of respondents in different modes: the features of HHs and municipalities identified in this study provide useful information in order to develop specific campaign and support for the web-based questionnaire. Knowledge about which sample members are more or less likely to respond in which mode allows targeting particular survey mode strategies at specific subgroups, providing significant advantage also for successive waves of the survey.

Further development of the study could focus on the employment of a multilevel logistic model in which HHs would be the first level units and the regions the second level units, since the geographical distribution of the attitude towards web surveys appeared so important. Moreover, the study of the stability of the profiles over time could be a useful improvement for the analysis.

References

1. Antoun, C., Couper, M.P., Conrad, F.G.: Effects of mobile versus PC web on survey response quality: a crossover experiment in a probability web panel. Public Opin. Q. **81**, 280–306 (2017)
2. Benassi, F., Naccarato, A.: HHs in potential economic distress. A geographically weighted regression model for Italy, 2001–2011. Spat. Stat. **21**, 362–376 (2017)
3. Bethlehem, I., Cobben, F., Schouten, B.: Handbook of Nonresponse in HH Surveys. Wiley, New Jersey, USA (2011)
4. Bianchi, A., Biffignandi, S., Lynn, P.: Web-face-to-face mixed-mode design in a longitudinal survey: effects on participation rate, sample composition, and costs. J. Official Stat. **33**(2), 385–408 (2017)
5. Biffignandi, S., Pratesi, M.: Modelling the respondents' profile in a web survey on firms in Italy. In: Ferligoj, A., Mrvar, A. (eds.) Development in Social Science Methodology. FDV, Metodoloski zvezki, Ljubljana (2002)
6. Blom, A.G., Bosnjak, M., Das, S., Cornilleau, A., Cousteaux, A., Douhou, S., Krieger, U.: A comparison of four probability-based online and mixed-mode panels in Europe. Soc. Sci. Comput. Rev. 1–18 (2015)
7. Calinescu, M., Schouten, B.: Adaptive survey designs to minimize survey mode effects—a case study on the Dutch Labor Force Survey. Surv. Methodol. **41**(2), 403–425 (2015)
8. Cellini, R., Torrisi, G.: Regional resilience in Italy: a very long-run analysis. Reg. Stud. **48**(11), 1779–1796 (2014)
9. Chieppa, A., Gallo, G., Tomeo, V., Borrelli, F., di Domenico, S.: Knowledge discovery for inferring the usually resident population from administrative registers. Math. Popul. Stud. **26**(2), 96–102 (2018)
10. Citro, C.F.: From multiple modes for surveys to multiple data sources for estimates. Surv. Methodol. **40**(2), 137–161 (2014)
11. Cobben, F., Bethlehem, J.G.: Web panels for official statistics, Discussion paper 201307. Statistics, The Hague, The Netherlands (2013)

12. Cracolici, M.F., Cuffaro, M., Nijkamp, P.: Geographical distribution of unemployment: an analysis of provincial differences in Italy. Growth Chang. **38**(4), 649–670 (2007)
13. Crescenzi, F., Sindoni, G.: The combined use of multiple data sources in the population census. In: Proceedings of the UNECE Group of Experts on Population and Housing Censuses, 30 September–2 October 2015, Geneva (2015)
14. D'Alò, M., Falorsi, S., Fasulo, A., Solari, F.: Sample design for the integration of population census and social surveys. In: Petrucci, A., Racioppi, F., Verde, R. (eds.) New Statistical Developments in Data Science. SIS 2017, Florence, Italy, June 28–30, pp. 191–202. Springer International Publishing, Switzerland (2019)
15. Durrant, G.B., Steele, F.: Multilevel modelling of refusal and non-contact in HH surveys: evidence from six UK government surveys. J. R. Stat. Soc. Ser. A—Stat. Soc. 361–381 (2009)
16. Espinosa, J., Hennig, C.: A constrained regression model for an ordinal response with ordinal predictors. Stat. Comput. **29**, 869–890 (2019)
17. European Union: Regulation of the European Parliament as Regards the Territorial Typologies 2017/2391. Official Journal of the European Union (2017)
18. Eurostat: Digital Economy and Society Statistics—HHs and Individuals. Luxemburg (2020). https://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_HHs_and_individuals
19. Eurostat: Applying the Degree of Urbanisation. A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons, 2021 edn. Luxemburg, Eurostat (2021)
20. Gallo, G., Paluzzi, E.: I censimenti nell'Italia unita. Le fonti di stato della popolazione tra il XIX e il XXI secolo. In: "I censimenti fra passato, presente e futuro" Torino, 4–6 dicembre 2010, ANNALI DI STATISTICA ANNO 141—SERIE XII—VOL. 2, ISTAT (2012)
21. Hargittai, E.: Second-level digital divide: differences in people's online skills. First Monday **7**(4) (2002)
22. ISTAT: Preliminary experimental results on the Italian population and housing census estimation methods. In: Proceedings of the Twentieth Conference of European Statisticians Group of Experts on Population and Housing Censuses, 26–28 September, 2018. United Nations Economic Commission for Europe, Geneva (2018)
23. ISTAT: Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età, cittadinanza, grado di istruzione e condizione professionale per gli anni 2018 e 2019 (2020)
24. Jensen, P.: Towards a register based statistical system—some Danish experience. Stat. J. **1**(3), 341–365 (1983)
25. de Leeuw, E.D.: Mixed mode: past, present, and future. Surv. Res. Methods **12**(2), 75–89 (2018)
26. de Leeuw, E.D., Suzer-Gurtekin, Z.T., Hox, J.J.: The design and implementation of mixed-mode surveys. In: Johnson, P., Pennell, B.E., Stoop, I.A.L., Dorer, B. (eds.) Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural contexts (3MC), pp. 387–408. Wiley, New York (2019)
27. Maslovskaya, O., Durrant, G.B., Smith, P.W.F., Hanson, T., Villar, A.: What are the characteristics of respondents using different devices in mixed-device online surveys? Evidence from six UK surveys. Int. Stat. Rev. **87**(2), 326–346 (2019)
28. Pratesi, M., Lozar Manfreda, K., Biffignandi, S., Vehovar, V.: List-based web surveys: quality, timeliness and nonresponse in the steps of the participation flow. J. Official Stat. (2004)
29. Rivero, M.S., Rangel, M.C.R., Martín, J.M.S.: Geotourist profile identification using binary logit modeling: application to the Villuercas-Ibores-Jara Geopark (Spain). Geoheritage **11**, 1399–1412 (2019)
30. UNECE (United Nations Economic Commission for Europe): Population definitions at the 2010 censuses round in the countries of the UNECE region. Paper presented to the Fifteenth Meeting of Group of Experts on Population and Housing Censuses. Geneva, 30 September–3 October 2013 (2013)
31. UNECE (United Nations Economic Commission for Europe): Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses. ECE/CES/STAT/2018/4. United Nations, New York, UNECE (2018)

32. UNSD (United Nations Statistical Division): Guidelines on the Use of Electronic Data Collection Technologies in Population and Housing Censuses. United Nations, New York (2019)
33. Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Boston (2012)

Publisher Correction to: Forecasting Combination of Hierarchical Time Series: A Novel Method with an Application to CoVid-19

Livio Fenga

Publisher Correction to:
Chapter "Forecasting Combination of Hierarchical Time Series: A Novel Method with an Application to CoVid-19" in: N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_14

The original version of the chapter was inadvertently published with incorrect coauthor names "Fulvio De Santis and Stefania Gubbiotti" which have now been removed and Livio Fenga is the only author for Chapter 14.

The correction chapter and the book have been updated with the changes.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-031-16609-9_14

N. Salvati et al. (eds.), *Studies in Theoretical and Applied Statistics*, Springer Proceedings in Mathematics & Statistics 406, https://doi.org/10.1007/978-3-031-16609-9_33