**Social Indicators Research manuscript No.**
(will be inserted by the editor)

# An Italian Composite Subjective Well-being Index
## the Voice of Twitter Users from 2012 to 2017

**S. M. Iacus · G. Porro · S. Salini · E. Siletti**

Received: date / Accepted: date

**Abstract** Since 2012, driven by the desire to propose a subjective well-being index complementary to the traditional measures, with high time and space frequency, our team evaluates, analysing Twitter data, a composite index that captures various aspects and dimensions of individual and collective life. The Subjective Well-being Index (SWBI) is a multidimensional indicator whose components were inspired by the dimensions adopted for the Happy Planet Index provided by the New Economic Foundation. In detail, it consists of eight dimensions that describe three different areas: personal well-being, social well-being and well-being at work.

The Italian Subjective Well-being Index (SWBI$_{ITA}$), that we display here, audits the Italian subjective well-being revealed by tweets acquired via the public Twitter API, written in the Italian language, and posted from Italy from January 2012 to December 2017. Around 1 to 5% of the data includes geo-referenced information, which allows us to provide an index at local level. The Twitter data analysis is carried on with a human supervised sentiment analysis method, the Integrated Sentiment Analysis (iSA) algorithm. In this work, after a weighting procedure adopted to partially overcome the selection bias caused by the use of data from social network, we describe the SWBI$_{ITA}$ dimensions in the considered period at the regional level. Moreover, for some dimensions, for which a similar currently available measure provided by Italian official statistics exists, comparisons are proposed emphasizing novelties, similarities and differences.

**Keywords** Composite indicators · Quality of life · Sentiment analysis · Big data · Social network

**Stefano Maria Iacus** (ORCID: 0000-0002-4884-0047)
**Silvia Salini** (ORCID: 0000-0001-6106-9835)
**Elena Siletti** (Corresponding author; ORCID: 0000-0003-0376-5992)
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano,
Via Conservatorio, 7 - 20122 Milano
E-mail: stefano.iacus@unimi.it, silvia.salini@unimi.it, elena.siletti@unimi.it

**Giuseppe Porro** (ORCID: 0000-0001-6270-0377)
Department of Law, Economics and Culture, Università degli Studi dell'Insubria,
Via Sant'Abbondio, 12 - 22100 Como
E-mail: giuseppe.porro@uninsubria.it

**Social Indicators Research manuscript No.**
(will be inserted by the editor)

# An Italian Composite Subjective Well-being Index

## the Voice of Twitter Users from 2012 to 2017

**Author1 · Author2 · Author3 · Author4**

Received: date / Accepted: date

**Abstract** Since 2012, driven by the desire to propose a subjective well-being index complementary to the traditional measures, with high time and space frequency, our team evaluates, analysing Twitter data, a composite index that captures various aspects and dimensions of individual and collective life. The Subjective Well-being Index (SWBI) is a multidimensional indicator whose components were inspired by the dimensions adopted for the Happy Planet Index provided by the New Economic Foundation. In detail, it consists of eight dimensions that describe three different areas: personal well-being, social well-being and well-being at work.

The Italian Subjective Well-being Index (SWBI$_{ITA}$), that we display here, audits the Italian subjective well-being revealed by tweets acquired via the public Twitter API, written in the Italian language, and posted from Italy from January 2012 to December 2017. Around 1 to 5% of the data includes geo-referenced information, which allows us to provide an index at local level. The Twitter data analysis is carried on with a human supervised sentiment analysis method, the Integrated Sentiment Analysis (iSA) algorithm. In this work, after a weighting procedure adopted to partially overcome the selection bias caused by the use of data from social network, we describe the SWBI$_{ITA}$ dimensions in the considered period at the regional level. Moreover, for some dimensions, for which a similar currently available measure provided by Italian official statistics exists, comparisons are proposed emphasizing novelties, similarities and differences.

**Keywords** Composite indicators · Quality of life · Sentiment analysis · Big data · Social network

## 1 Introduction

In the last ten years, the approach to evaluate well-being evolved as never before. Scholars, driven by the Stiglitz Commission (Stiglitz et al., 2009), proposed a great number of indices with different structures and considering a great variety of dimensions. With respect to the past, the more innovative suggestion coming from these new proposals is the importance related to the subjective dimension of well-being, that now it is considered essential like the more traditional quantitative dimensions (as economic indicators, such as Gross Domestic Product). Considering the traditional or official methodologies, the obtaining of subjective and perceived information of this type is closely linked to the use of sample surveys that, despite all the attempts made (Schwarz, 1999; Schwarz and Strack, 1999; Kahneman and Krueger, 2006), still show some methodological drawbacks (Deaton, 2012; Feddersen et al., 2016). In short summary, these weak points have to do with explicit questions that can be influenced by contextual elements and the suffering of non-response bias. Moreover, conventional surveys are expensive and make it almost impossible to obtain high-frequency data or with an adequate territorial granularity. An objective evaluation of the indices currently proposed by the official statistics cannot fail to note a limited and undersized presence of the subjective and perceived dimension of well-being (Iacus et al., 2017). Trying to fill this information gap, our team, with the aim of finding complementary indicators to the traditional ones, has tried to obtain an index of subjective well-being by using big data. In particular, we focused on the use of data coming from social networks: as far as this work is concerned, we make use of information from Twitter. The paper is organized as follows: in Sect.2 we summarize the SWBI features as a multidimensional subjective well-being index. In Sect.3 the index for the Italian regions, with a quarterly frequency from 2012 to 2017, is illustrated. In Sect.4 a comparison between the traditional available measures of subjective well-being and the SWBI$_{ITA}$ is discussed. Finally our conclusions can be found in Sect.5.

## 2 The Subjective Well-being Index - SWBI

Social media has been recently considered a huge source of information or big data, in particular, Kwong et al. (2012) and Hofacker et al. (2016) defined them as the largest available focus group in the world. The advantages related to deal with these new data are: they cover all kind of topics, are continuously updated, involve different people, are (mostly) free from censorship and, last but not least, come to social analysts in huge amount for free or with little costs. On the other hand, this kind of information is not free of disadvantages. To be a social media user, one has to overcome some barriers: to have Internet access, to open a social media account, and to actively use it [1]. Even if the number of social media users is always increasing, not all the members of a community are social media users; hence, one of the main issue linked to the use of big data coming from social networks concerns sampling bias. Since it is clearly not possible to renounce such an intense source of information, scholars are still working to the solution for these issues. In particular, Iacus et al. (forthcoming) recently proposed to mash-up official statistics with Twitter data to consider sampling bias.

Since 2012, we propose to use a human supervised machine learning method (Integrated Sentiment Analysis *iSA*, (Ceron et al., 2016)) on Twitter data to obtain a composite index of subjective and perceived well-being that captures various aspects and dimensions of individual and collective life (Curini et al., 2015; Iacus et al., 2015).

[1] Penetration data from the We Are Social and Hootsuite's report ("Digital in 2019", Jan 2019; available at http://wearesocial.com): looking at the world population 57% (+9.1%) has Internet access, and 45% (+9%) has a social media account and makes an active use of it, while in Italy these percentages are respectively 92% (+27%) and 59%(+2.9%). Annual digital growth from January 2018 to January 2019 in brackets.

**Table 1** The SWBI structure

| *Personal well-being* |
|---|
| (`emo`) **emotional well-being**: the overall balance between the frequency of experiencing positive and negative emotions, with higher scores showing that positive feelings are felt more often than negative ones;<br>(`sat`) **satisfying life**: having a positive assessment of one's life overall;<br>(`vit`) **vitality**: having energy, feeling well-rested and healthy while also being physically active;<br>(`res`) **resilience and self-esteem**: a measure of individual psychological resources, of optimism and of the ability to deal with life stress;<br>(`fun`) **positive functioning**: feeling free to choose and having the opportunity to do it; being able to make use of personal skills while feeling absorbed and gratified in daily activities; |
| *Social well-being* |
| (`tru`) **trust and belonging**: trusting other people, feeling treated fairly and respectfully while experiencing sentiments of belonging;<br>(`rel`) **relationships**: the degree and quality of interactions in close relationships with family, friends and others who provide support; |
| *Well-being at work* |
| (`wor`) **quality of job**: feeling satisfied with a job, experiencing satisfaction with work-life balance, evaluating the emotional experiences of work and work conditions. |

The SWBI (Subjective Well-being Index) index is a multidimensional well-being indicator whose components were inspired by the Happy Planet Index dimensions proposed by the New Economic Foundation (New Economics Foundation, 2012). It consists of eight dimensions that concern three different well-being areas: personal well-being, social well-being and well-being at work (Table 1).

With *iSA* a sample of texts is read by human coders, during the preparation of the training set, and then the rest of the corpus is classified by the algorithm. Each tweet is classified according to the scale -1, 0, 1, where -1 is negative, 0 is neutral and 1 is positive feeling.

For example, a text like "I am grateful to my friends and relatives who sustained me during my hard times" is classified as $\texttt{rel} = +\texttt{1}$. While a text like "you can't really trust anyone nowadays" is classified as $\texttt{tru} = -\texttt{1}$; or a text like "ok, let's go to work again today" as $\texttt{wor} = \texttt{0}$. Some adopted coding rules:

- off-topic texts are marked appropriately;
- if the readers are not fully convinced about the semantic context of the tweet: they do not classify the text, they skip it and classify another one;
- only self-expressed or individual expression of well-being or own views of the tweeter are considered;
- re-tweet are ok, because the tweeter shares the same view;
- each tweet can be classified along one or more dimensions.

The major feature of this analysis is that it does not rely on dictionaries or special semantic rules.

It should be noted, as a specific characteristic, that SWBI is not the result of the aggregation of people perceived well-being measurements, but it directly estimates the aggregate composition of the mood within the society. The SWBI index has been in place for Italy and Japan since 2012.

## 3 The Italian Subjective Well-being Index - SWBI$_{ITA}$ from 2012 to 2017

The SWBI$_{ITA}$'s data source are tweets written in Italian, posted from Italy, and acquired via the public Twitter API. Data over the 24 quarters from 2012 to 2017 were considered at the provincial and regional levels, thanks to the geo-reference information, and more than 240 million of tweets were downloaded and classified. The variability of the number of tweets is remarkable, both along the time and the space dimension, as the range of data extends from a minimum of 1727 tweets in 2016-Q1 for the Basilicata region to a maximum of 2,728,640 in 2017-Q2 for the Lombardia region. It is important to notice that, in relation to the population, the use of Twitter is homogeneous in all the Italian provinces (Iacus et al., 2017).

Traditional methods adopted in the literature to face the sampling bias problem when using non-representative samples (e.g., the propensity score weighting (Rosembaum and Rubin, 1983) or the Heckman correction (Heckman, 1979)) are based on the use of unit level data (Cooper and Greenaway, 2015). In the case of the SWBI the information about every Twitter user are not get-at-able and only aggregated data are available. For this reason, to deal with sampling bias, we partially follow Iacus et al. (forthcoming), performing a weighted method. Unfortunately, the second step of the proposal, that require the use of official statistics with high time frequencies and an Italian provinces granularity in Small Area Models (SAE), is not practicable for all the SWBI$_{ITA}$ indicators due to the lack of necessary official statistics.

Especially, we adopt a hierarchical aggregation, where province level data, weighted by the characteristics of provincial macro-variables, are used to estimate the composition of sentiment throughout regional society. For each indicator, we apply $\hat{y}^w_{dt}$ as the regional sampling mean, where the regional units are the weighted means of province level units, in order to overcome the non-random sampling structure of the data:

$$\hat{y}^w_{dt} = \frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt} w_{idt}, \tag{1}$$

where $n_{dt}$ is the number of provinces in region $d$ at time $t$, and $w_{idt}$ are the weights used [2].

To implement this weighting procedure, we use as weights the Twitter rate and the broadband coverage. The Twitter rate, the ratio between the number of tweets analysed and the population size in the region in the same period, computed in each period and at province level, can be considered as a good proxy of the use of Twitter for Italians. The average Twitter rate is around 18% ($SD = 12.29$), with a minimum regional value higher than 9% ($SD = 4.93$) in Campania, and a maximum regional value higher than 30% ($SD = 21.15$) in Molise. The spread during the observational period is lower for large regions like Lazio, Puglia, Campania and Lombardia, while being higher for small regions like Molise and Marche. The broadband coverage is annual public data provided by *Il Sole 24 Ore* and *Infratel Italia* for all the Italian provinces and can be considered the opportunity to access

[2] Notice that Valle d'Aosta region has been dropped from the analysis because, considering that it consists of a single province, the proposed approach is not applicable.

the Internet in different provinces. Coverage is quite stationary during a single year but, over time, what can happen is only an improvement of coverage in space or in signal intensity. Therefore, we replace the missing values with the data from the previous year to ensure that the coverage is not overestimated. The average broadband coverage is around 94% ($SD = 4.68$), with a minimum regional value of 72% ($SD = 4.57$) for Isernia in the Molise region. In 2012, the coverage mean was 92.15% ($SD = 3.9$) and in 2017, it was 92.65% ($SD = 5.6$). So, during the examined time period, the average broadband coverage remained quite the same, but the variability among regions increased, with an increment of around 42%.

In detail, calling $w_{1,idt}$ the Twitter rate and $w_{2,idt}$ the broadband coverage, to apply to weighting procedure for $\hat{y}^w_{dt}$ in (1), we computed the weights as $w_{idt} = w_{1,idt} \cdot w_{2,idt}$.

### 3.1 The Voice of Italian Twitter Users from 2012 to 2017

In this section we describe the eight well-being dimensions, that represent the voice of Italian Twitter users in the 24 quarters from 2012 to 2017. We comment on each dimension both in time, for each Italian region (for synthesis in the text Fig. 1 and Fig. 2, other figures in Appendix A, and Tab. 2, 3), where the units are the values of the index in the 24 quarters; and in space, for each quarter from 2012 to 2017 (for synthesis in the text Fig. 3 and Fig. 4, other figures in Appendix A, and Tab. 4, 5), where the units are the values of the index in the different regions.

All the dimensions present a very low variability between the regions (with coefficients of variations in Tab. 3: minimum for `rel` in Q2-2016 (0.0098), maximum for `rel` in Q2-2016 (0.3453), mean = 0.0546), that on average is a fifth of the variability over time (with coefficients of variations in Tab. 5: minimum for Lazio (0.0504), maximum for Umbria (0.7727), mean = 0.2641). Just looking at the variability in time we notice that the dimension less variable for all the regions (except Basilicata) is the resilience and self-esteem (`res`), while the most variable is always the quality of job (`wor`). The great variability over time is often due

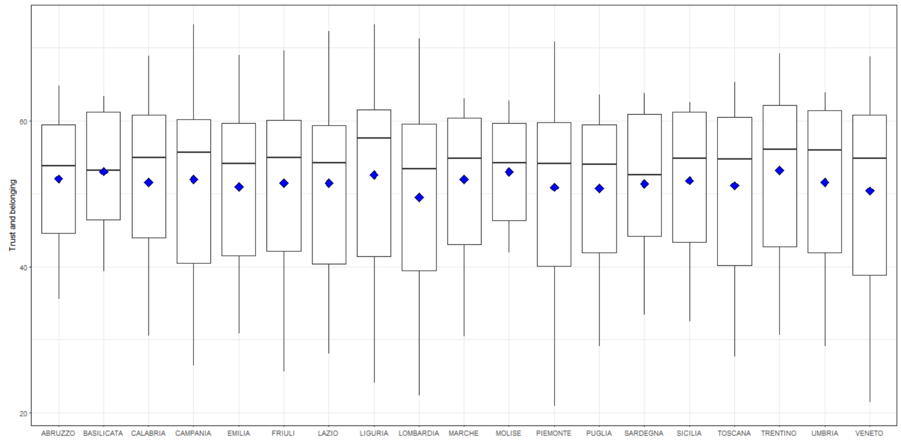

**Fig. 1** Distribution of trust and belonging (`tru`) for the Italian regions

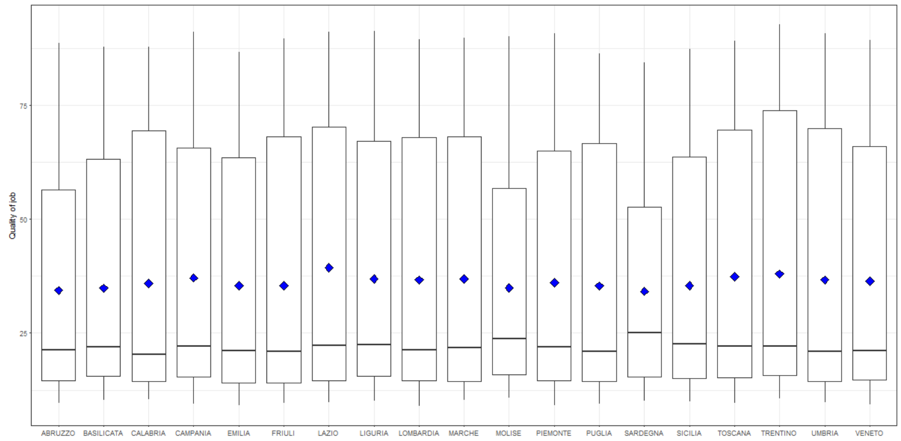

**Fig. 2** Distribution of quality of job (`wor`) for the Italian regions

to real peaks that are visible in the box plots with long whiskers or identified as anomalous values, sometimes both on the lower and upper tails of the distributions (e.g., `sat` or `fun`).

Looking at the means, the quality of job (`wor`) is the dimension with the lowest values for all the regions, while positive functioning (`fun`) has always the highest value. In Q2-2012 for (`wor`) we observe the minimum value of the mean (9.88) and in Q3-2016 for vitality (`vit`) the highest value for the mean (91.88).

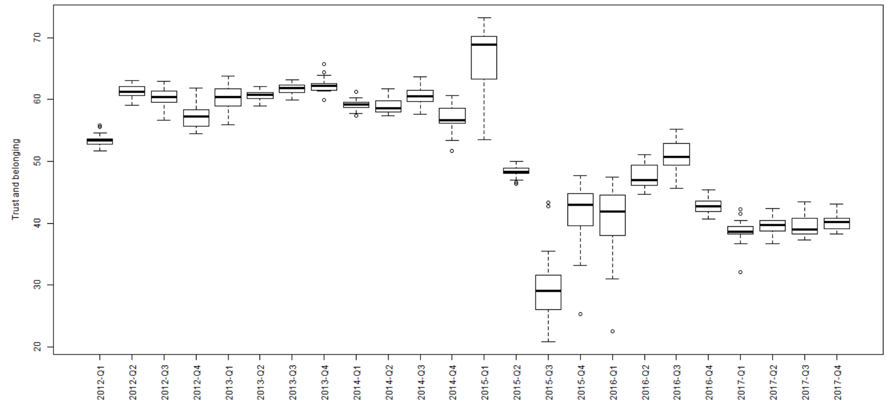

**Fig. 3** Distribution of trust and belonging (`tru`) in the 24 quarters from 2012 to 2017

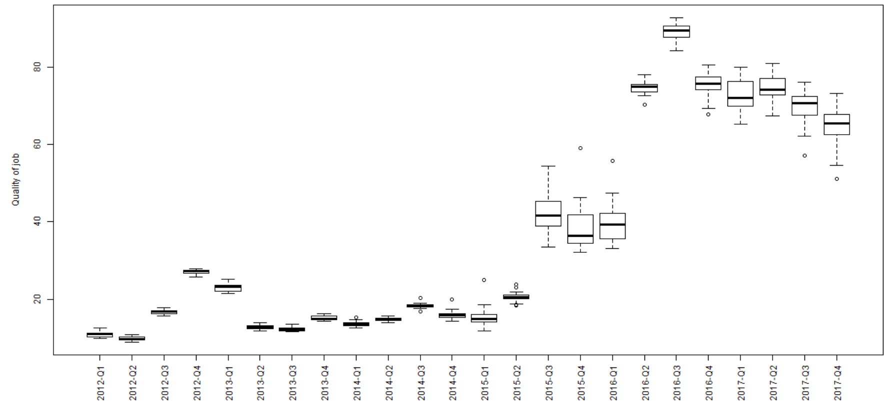

**Fig. 4** Distribution of quality of job (`wor`) in the 24 quarters from 2012 to 2017

To confirm this structure we perform a within subject (region) one-way ANOVA (ANalysis Of VAriance) for each SWBI dimension. Results in Tab. 6 confirm that for all the indicators time effect is the most important.

Focusing on the time trend, quite all the dimensions have a similar trend as quality of job (`wor` in Tab. 4), with low and very similar values up to 2015 with an evident increase and much higher values in the following periods. While for trust and belonging (`tru` in Tab. 3), the trend is different, with higher values in the first

**Table 2** Average values for the eight indicators in the Italian region, in the 24 quarters from 2012 to 2017

| Regions | emo | fun | rel | res | sat | tru | vit | wor |
|---|---|---|---|---|---|---|---|---|
| Abruzzo | 50.54 | 63.94 | 52.20 | 52.14 | 46.90 | 52.03 | 58.06 | 34.38 |
| Basilicata | 51.12 | 65.22 | 52.79 | 53.21 | 48.34 | 53.02 | 58.88 | 34.87 |
| Calabria | 48.69 | 64.88 | 52.07 | 52.58 | 48.56 | 51.58 | 60.00 | 35.96 |
| Campania | 51.53 | 65.51 | 52.89 | 52.46 | 49.02 | 51.96 | 62.09 | 37.11 |
| Emilia | 52.49 | 64.61 | 52.98 | 52.08 | 45.94 | 50.96 | 60.08 | 35.43 |
| Friuli | 50.34 | 64.08 | 52.11 | 51.71 | 47.46 | 51.44 | 61.77 | 35.38 |
| Lazio | 50.25 | 65.54 | 53.12 | 52.78 | 48.21 | 51.47 | 59.82 | 39.31 |
| Liguria | 49.07 | 65.69 | 51.61 | 52.76 | 50.09 | 52.58 | 61.67 | 36.87 |
| Lombardia | 53.13 | 63.17 | 52.41 | 52.30 | 46.96 | 49.48 | 61.94 | 36.68 |
| Marche | 51.58 | 65.02 | 53.12 | 52.63 | 46.58 | 51.98 | 58.35 | 36.86 |
| Molise | 49.64 | 64.25 | 51.96 | 53.31 | 49.24 | 53.01 | 58.23 | 34.94 |
| Piemonte | 51.57 | 64.10 | 52.17 | 52.30 | 47.49 | 50.87 | 61.29 | 36.11 |
| Puglia | 51.88 | 63.76 | 53.21 | 51.99 | 46.20 | 50.73 | 60.10 | 35.34 |
| Sardegna | 51.86 | 63.54 | 52.88 | 52.38 | 47.41 | 51.36 | 58.73 | 34.16 |
| Sicilia | 50.18 | 64.10 | 51.94 | 52.19 | 47.64 | 51.80 | 59.39 | 35.37 |
| Toscana | 51.70 | 65.21 | 53.07 | 52.48 | 47.84 | 51.10 | 60.72 | 37.41 |
| Trentino | 50.11 | 67.84 | 52.00 | 53.49 | 51.45 | 53.19 | 62.92 | 37.99 |
| Umbria | 50.19 | 65.51 | 52.61 | 52.90 | 48.65 | 51.53 | 60.46 | 36.66 |
| Veneto | 52.09 | 64.37 | 52.49 | 52.22 | 47.28 | 50.39 | 61.04 | 36.43 |

**Table 3** Coefficients of variation for the eight indicators in the Italian regions, in the 24 quarters from 2012 to 2017

| Regions | emo | fun | rel | res | sat | tru | vit | wor |
|---|---|---|---|---|---|---|---|---|
| Abruzzo | 0.1116 | 0.1612 | 0.2981 | 0.0691 | 0.2409 | 0.1660 | 0.2033 | 0.7418 |
| Basilicata | 0.0680 | 0.1550 | 0.2973 | 0.0727 | 0.1944 | 0.1623 | 0.2119 | 0.7367 |
| Calabria | 0.1380 | 0.1747 | 0.3041 | 0.0815 | 0.2475 | 0.2059 | 0.2024 | 0.7660 |
| Campania | 0.1439 | 0.1907 | 0.3161 | 0.0979 | 0.2930 | 0.2171 | 0.1823 | 0.7467 |
| Emilia | 0.1380 | 0.1552 | 0.3038 | 0.0569 | 0.2643 | 0.1999 | 0.1900 | 0.7406 |
| Friuli | 0.1516 | 0.1967 | 0.3209 | 0.1114 | 0.2831 | 0.2061 | 0.1840 | 0.7635 |
| Lazio | 0.1536 | 0.1565 | 0.3225 | 0.0504 | 0.2709 | 0.2124 | 0.2267 | 0.7438 |
| Liguria | 0.1663 | 0.1974 | 0.3394 | 0.1009 | 0.2929 | 0.2309 | 0.1831 | 0.7419 |
| Lombardia | 0.1636 | 0.2382 | 0.3126 | 0.0715 | 0.2888 | 0.2685 | 0.1849 | 0.7485 |
| Marche | 0.0999 | 0.1597 | 0.3230 | 0.0609 | 0.2640 | 0.1817 | 0.2403 | 0.7462 |
| Molise | 0.1036 | 0.1589 | 0.3023 | 0.0691 | 0.2279 | 0.1385 | 0.2018 | 0.7269 |
| Piemonte | 0.1559 | 0.1969 | 0.3147 | 0.0609 | 0.2845 | 0.2348 | 0.1889 | 0.7455 |
| Puglia | 0.1221 | 0.1745 | 0.2923 | 0.0807 | 0.2581 | 0.1961 | 0.1911 | 0.7432 |
| Sardegna | 0.0950 | 0.1602 | 0.2810 | 0.0868 | 0.2197 | 0.1834 | 0.1789 | 0.6707 |
| Sicilia | 0.1035 | 0.1692 | 0.3054 | 0.0795 | 0.2386 | 0.1885 | 0.1957 | 0.7324 |
| Toscana | 0.1132 | 0.1817 | 0.3146 | 0.0766 | 0.2679 | 0.2132 | 0.2089 | 0.7354 |
| Trentino | 0.1199 | 0.1863 | 0.3422 | 0.0964 | 0.2619 | 0.2093 | 0.2208 | 0.7545 |
| Umbria | 0.1154 | 0.1957 | 0.3127 | 0.0946 | 0.2809 | 0.2101 | 0.2187 | 0.7727 |
| Veneto | 0.1380 | 0.1938 | 0.3130 | 0.0698 | 0.2770 | 0.2466 | 0.1851 | 0.7398 |

quarters, a decrease in mid-2015, then a slight recovery that did not, however, make it possible to return to previous levels.

**Table 4** Average values for the eight indicators in the 24 quarters from 2012 to 2017

| Quarter | emo | fun | rel | res | sat | tru | vit | wor |
|---|---|---|---|---|---|---|---|---|
| Q1-2012 | 58.79 | 64.97 | 36.94 | 58.09 | 41.74 | 53.42 | 52.55 | 10.94 |
| Q2-2012 | 55.21 | 64.79 | 38.38 | 54.82 | 37.88 | 61.38 | 51.04 | 9.88 |
| Q3-2012 | 56.99 | 65.06 | 36.73 | 53.34 | 43.51 | 60.51 | 53.49 | 16.79 |
| Q4-2012 | 56.11 | 68.91 | 34.68 | 53.98 | 51.68 | 57.62 | 53.25 | 27.18 |
| Q1-2013 | 55.14 | 71.05 | 36.21 | 56.58 | 53.28 | 60.73 | 52.10 | 23.18 |
| Q2-2013 | 52.74 | 71.19 | 41.53 | 53.29 | 57.37 | 60.71 | 60.39 | 13.00 |
| Q3-2013 | 56.08 | 70.59 | 41.23 | 57.67 | 53.14 | 61.94 | 56.39 | 12.55 |
| Q4-2013 | 53.69 | 69.52 | 41.27 | 57.61 | 49.23 | 62.32 | 55.06 | 15.37 |
| Q1-2014 | 53.53 | 68.92 | 42.56 | 57.04 | 51.31 | 59.33 | 54.54 | 13.96 |
| Q2-2014 | 48.19 | 66.94 | 45.27 | 54.53 | 49.41 | 59.23 | 54.90 | 15.10 |
| Q3-2014 | 49.53 | 66.04 | 44.78 | 54.01 | 47.09 | 60.86 | 51.45 | 18.53 |
| Q4-2014 | 43.43 | 58.02 | 39.55 | 55.64 | 51.05 | 56.79 | 52.42 | 16.17 |
| Q1-2015 | 35.09 | 61.77 | 42.55 | 52.70 | 50.85 | 66.41 | 54.20 | 15.75 |
| Q2-2015 | 44.59 | 52.65 | 45.33 | 52.23 | 44.36 | 48.41 | 50.97 | 20.79 |
| Q3-2015 | 62.26 | 45.29 | 60.89 | 43.79 | 32.91 | 29.64 | 56.42 | 42.07 |
| Q4-2015 | 51.18 | 37.22 | 57.01 | 43.84 | 18.19 | 41.23 | 49.93 | 38.70 |
| Q1-2016 | 53.56 | 37.03 | 52.91 | 50.24 | 19.95 | 40.19 | 51.76 | 39.70 |
| Q2-2016 | 46.57 | 71.65 | 78.84 | 47.37 | 62.45 | 47.71 | 82.92 | 74.74 |
| Q3-2016 | 42.19 | 85.90 | 91.56 | 50.40 | 77.36 | 50.93 | 91.88 | 89.43 |
| Q4-2016 | 46.51 | 75.17 | 77.63 | 50.15 | 59.37 | 43.16 | 78.12 | 74.74 |
| Q1-2017 | 47.67 | 71.24 | 66.89 | 50.24 | 51.53 | 38.68 | 71.57 | 72.91 |
| Q2-2017 | 49.20 | 72.18 | 69.61 | 51.44 | 53.12 | 39.63 | 74.67 | 74.40 |
| Q3-2017 | 48.28 | 71.46 | 68.47 | 53.01 | 52.78 | 39.67 | 70.59 | 69.65 |
| Q4-2017 | 50.11 | 67.59 | 66.95 | 50.90 | 48.71 | 40.23 | 68.14 | 64.10 |

**Table 5** Coefficients of variation for the eight indicators in the 24 quarters from 2012 to 2017

| Quarter | emo | fun | rel | res | sat | tru | vit | wor |
|---|---|---|---|---|---|---|---|---|
| Q1-2012 | 0.0427 | 0.0176 | 0.0415 | 0.0229 | 0.0418 | 0.0193 | 0.0206 | 0.0621 |
| Q2-2012 | 0.0324 | 0.0099 | 0.0233 | 0.0112 | 0.0571 | 0.0178 | 0.0142 | 0.0552 |
| Q3-2012 | 0.0305 | 0.0266 | 0.0358 | 0.0293 | 0.0613 | 0.0283 | 0.0334 | 0.0359 |
| Q4-2012 | 0.0398 | 0.0521 | 0.0815 | 0.0515 | 0.1054 | 0.0461 | 0.0657 | 0.0202 |
| Q1-2013 | 0.0427 | 0.0436 | 0.0757 | 0.0509 | 0.0900 | 0.0408 | 0.0497 | 0.0489 |
| Q2-2013 | 0.0891 | 0.0225 | 0.0438 | 0.0291 | 0.0685 | 0.0160 | 0.0357 | 0.0590 |
| Q3-2013 | 0.0597 | 0.0164 | 0.0298 | 0.0317 | 0.0501 | 0.0140 | 0.0287 | 0.0890 |
| Q4-2013 | 0.0611 | 0.0204 | 0.0309 | 0.0318 | 0.0593 | 0.0202 | 0.0331 | 0.0809 |
| Q1-2014 | 0.0697 | 0.0210 | 0.0300 | 0.0363 | 0.0813 | 0.0188 | 0.0365 | 0.0996 |
| Q2-2014 | 0.0767 | 0.0240 | 0.0381 | 0.0318 | 0.0823 | 0.0282 | 0.0375 | 0.0650 |
| Q3-2014 | 0.0641 | 0.0257 | 0.0334 | 0.0364 | 0.0929 | 0.0331 | 0.0359 | 0.0714 |
| Q4-2014 | 0.0809 | 0.0339 | 0.0605 | 0.0201 | 0.0780 | 0.0406 | 0.0441 | 0.1001 |
| Q1-2015 | 0.1728 | 0.0660 | 0.0982 | 0.0183 | 0.1051 | 0.0776 | 0.0670 | 0.1782 |
| Q2-2015 | 0.0507 | 0.0282 | 0.0631 | 0.0158 | 0.0564 | 0.0213 | 0.0384 | 0.0677 |
| Q3-2015 | 0.0585 | 0.0651 | 0.0576 | 0.0484 | 0.0734 | 0.2030 | 0.0781 | 0.1191 |
| Q4-2015 | 0.0798 | 0.1493 | 0.0426 | 0.0787 | 0.3453 | 0.1316 | 0.0637 | 0.1649 |
| Q1-2016 | 0.0840 | 0.1503 | 0.0394 | 0.0851 | 0.2747 | 0.1492 | 0.0635 | 0.1359 |
| Q2-2016 | 0.1000 | 0.0222 | 0.0098 | 0.0282 | 0.0521 | 0.0439 | 0.0186 | 0.0241 |
| Q3-2016 | 0.1076 | 0.0250 | 0.0113 | 0.0205 | 0.0527 | 0.0587 | 0.0109 | 0.0281 |
| Q4-2016 | 0.0465 | 0.0369 | 0.0218 | 0.0123 | 0.0401 | 0.0435 | 0.0376 | 0.0577 |
| Q1-2017 | 0.0575 | 0.0397 | 0.0169 | 0.0190 | 0.0435 | 0.0549 | 0.0506 | 0.0535 |
| Q2-2017 | 0.0489 | 0.0336 | 0.0231 | 0.0196 | 0.0354 | 0.0396 | 0.0510 | 0.0520 |
| Q3-2017 | 0.0376 | 0.0430 | 0.0302 | 0.0186 | 0.0581 | 0.0472 | 0.0553 | 0.0697 |
| Q4-2017 | 0.0348 | 0.0448 | 0.0342 | 0.0231 | 0.0615 | 0.0339 | 0.0632 | 0.0862 |

**Table 6** Repeated measures ANOVA results (all p-value < 0.01)

| | | emo | | sat | |
|---|---|---|---|---|---|
| | Df | Sum Sq | F value | Sum Sq | F value |
| Area | 18 | 601.51 | | 793.71 | |
| Time | 23 | 15677.17 | 88.10 | 65258.75 | 299.04 |
| Residuals | 414 | 3203.18 | | 3928.06 | |
| | | fun | | tru | |
| Area | 18 | 480.28 | | 384.78 | |
| Time | 23 | 56945.73 | 395.64 | 46004.32 | 274.59 |
| Residuals | 414 | 2590.80 | | 3015.69 | |
| | | rel | | vit | |
| Area | 18 | 105.82 | | 897.37 | |
| Time | 23 | 115303.40 | 1225.04 | 61604.44 | 498.47 |
| Residuals | 414 | 1694.20 | | 2224.56 | |
| | | res | | wor | |
| Area | 18 | 94.04 | | 723.89 | |
| Time | 23 | 6399.60 | 88.79 | 312656.00 | 1778.45 |
| Residuals | 414 | 1297.32 | | 3164.44 | |

## 4 Traditional and new subjective well-being measures: a comparison

In this section we compare SWBI$_{ITA}$ dimensions with some indices provided by ISTAT in its "Aspects of daily life" report, because of their subjective nature. This survey is the same information source used for the subjective component of the BES (BES = "Benessere Equo e Sostenibile", i.e., in English, "Fair and Sustainable Well-being"), an index, proposed by ISTAT, after the Stiglitz's commission suggestions, to set up an equitable and sustainable well-being index. The ISTAT's sample survey "Aspects of daily life" collects fundamental details on Italian individual and household daily life. In this survey, there are several thematic areas on different social aspects useful to study the quality of life. Since 2005 this survey is annual with data collection in February. Yearly data are distributed free of charge, but they are representatives only for the five Italian geographical areas: North-west, North-east, Central, South, and Islands. To compare these data with SWBI$_{ITA}$ results, we aggregate SWBI yearly and according to the same geographical areas, weighting with the corresponding resident population.

In order to compare traditional and social network findings, we consider the following four ISTAT measures of satisfaction that seem to be defined in similarity with the SWBI dimensions:

- WS: the degree of *work satisfaction*, defined as the percentage of employed persons, aged 15 years and over, with a "good" level of satisfaction with their work. This index is computed as the sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. This index is compared with the quality of job (`wor`) for SWBI.
- SAT: the degree of *satisfaction for life*, defined as the percentage of persons, aged 14 years and over, with a "good" level of satisfaction for life in the complex. This index is computed as the sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. This index is compared with the satisfaction (`sat`) for SWBI.
- HEALTH: the degree of *health satisfaction*, defined as the percentage of persons, aged 14 years and over, with a "good" level of satisfaction for health. This index is computed as the sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. There is not a direct similarity with an SWBI dimension, but we try to compare it with the vitality (`vit`) dimension.
- RELS: the degree of *relational satisfaction*, defined as the percentage of persons, aged 14 years and over, with a "good" level of satisfaction with their relationship with family and friends. This index is the mean of two indices: the degree of satisfaction with the family members and the degree of satisfaction with friends. Both the initial indices are computed as sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. This index is compared with the degree and quality of interactions in close relationships with family, friends and others who provide support (`rel`) for SWBI.

All the correlations between the similar indices are displayed in Tab. 7.

**Table 7** Pearson correlation coefficients between some ISTAT "Daily life aspects" and SWBI$_{ITA}$ dimensions, in the five Italian geographical areas

| Area | Overall | North-west | North-east | Central | South | Islands |
|---|---|---|---|---|---|---|
| `wor` - WS | 0.245 | 0.694 | 0.383 | 0.581 | 0.849 | 0.480 |
| `sat` - SAT | 0.451 | 0.825 | 0.874 | 0.976 | 0.884 | 0.774 |
| `vit` - HEALTH | 0.106 | 0.462 | 0.253 | 0.821 | 0.858 | -0.445 |
| `rel` - RELS | -0.185 | -0.359 | -0.105 | -0.167 | -0.094 | -0.633 |

For the quality of job, if we consider the overall data, the correlation is about 25%, while if we analyse the relationships within each area we find stronger links, with a maximum value in South Italy equal to 85%.

Taking into account life satisfaction, if we consider the overall data, the correlation is about 45%, while if we analyse the relationships within each area we find stronger links, even more than those found for the quality of job, with a maximum value in Central Italy equal to 98%.

For vitality - health satisfaction, if we consider the overall data, the correlation is weaker and about 11%, while if we analyse the relationships within each area the links are stronger, but surprisingly for the Italian Islands the value is negative (-45%).

About the degree of relational satisfaction, all the correlations are negative, highlighting that the two dimensions considered have a reverse link.

Given the different scales of the ISTAT indices and the SWBI$_{ITA}$ dimensions, with the purpose of visual comparison, Figs. 5-8 represent the plots of their values both standardized. Looking at these plots, the correlations describe above are quite evident.

## 5 Conclusion

The new and rich amount of data provided by social media represents a revolution in data science. As can be expected, dealing with these data has both positive and negative aspects that need to be considered. However, it seems clear that ignoring

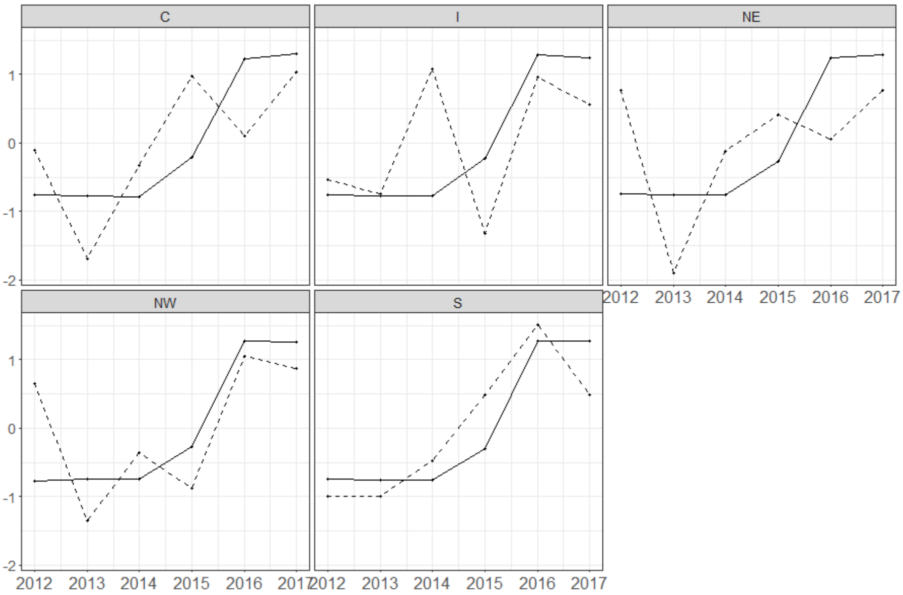

**Fig. 5** Standardized time series of SWBI-wor, solid line, and ISTAT's WS, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South)

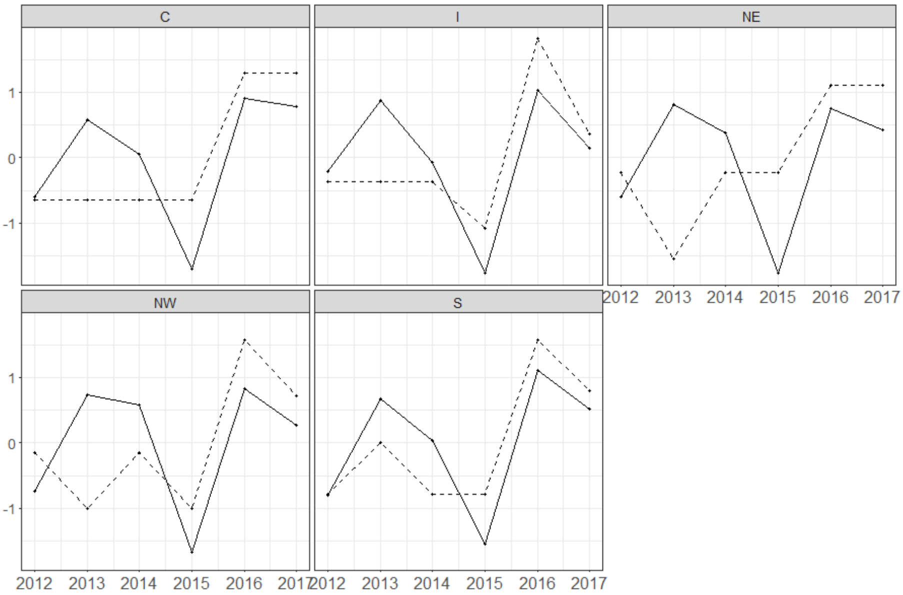

**Fig. 6** Standardized time series of SWBI-sat, solid line, and ISTAT's SAT, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South)

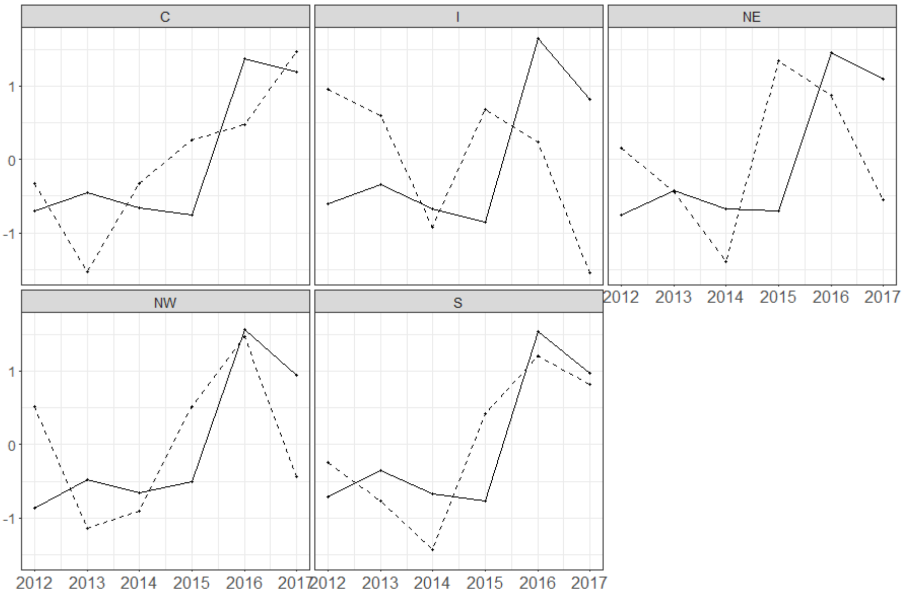

**Fig. 7** Standardized time series of SWBI-vit, solid line, and ISTAT's HEALTH, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South)

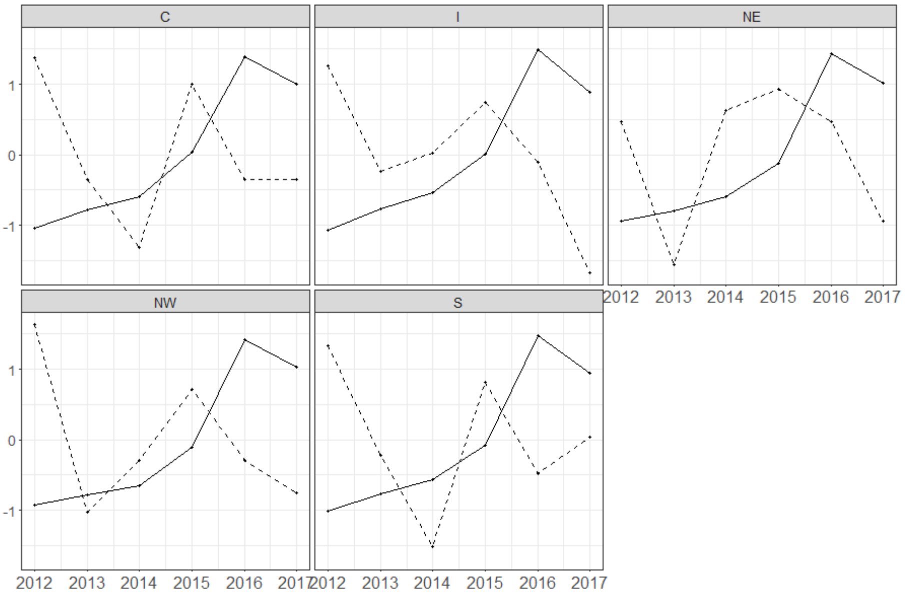

**Fig. 8** Standardized time series of SWBI-rel, solid line, and ISTAT RELS, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South)

them is not acceptable and that the best way to use these data is to link them to the official ones (Iacus et al., forthcoming).

In this paper, we show our experience with a new subjective and perceived well-being index, obtained mashing-up, with a weighting procedure, social network and official data. Weighting statistics based on social media, following Iacus et al. (forthcoming), we have corrected the selection bias up to the only benchmark data available, which are the official statistics.

To get this multidimensional well-being index we have applied a new human supervised sentiment analysis, which do not rely on dictionaries or special semantic rules, to Twitter Italian data.

We described time and space characteristics of the eight SWBI$_{ITA}$ dimensions, in 24 quarters from 2012 to 2017 for the Italian regions. A new peculiar finding is the great importance of time differences, that probably came to light thanks to the property of social data. Often with traditional well-being index, usually considered annually, we find a situation of temporal stationarity (see, e.g., ISTAT (2017)) that it is unrealistic in current life.

We have also shown that, despite using data from social network sites, some SWBI components (weighted `wor`-WS, `sat`-SAT and `vit`-HEALTH) correlate with the ISTAT statistics (available at macroeconomic level only) based on the traditional survey data.

It is worth emphasizing that - as for other variables based on social networks data, whose frequency is much higher, compared to more traditional statistics - absolute values of the index are not as relevant as the information it provides on trend and variability of phenomena.

This work is clearly just a starting point in the process that awaits us in the future, to deal with the social media data revolution. Integrating further big data sources with official information, each with its own bias corrected statistics, still remains a methodological challenges for future research.

**References**

Ceron A, Curini L, Iacus SM (2016) iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. Information Sciences 367-368:105–124, DOI https://doi.org/10.1016/j.ins.2016.05.052

Cooper D, Greenaway M (2015) Non-probability survey sampling in official statistics. Tech. rep., Office for National Statistics - Methodology Working Paper Series N.4

Curini L, Iacus S, Canova L (2015) Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. Social Indicators Research 121(2):525–542

Deaton A (2012) The Financial Crisis and the Well-Being of America, University of Chicago Press, Chicago, IL, chap 10, pp 343–368

Feddersen J, Metcalfe R, Wooden M (2016) Subjective wellbeing: why weather matters. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179(1):203–228

Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47(1):153–161

Hofacker CF, Malthouse EC, Sultan F (2016) Big data and consumer behavior: imminent opportunities. Journal of Consumer Marketing 33(2):89–97

Iacus SM, Porro G, Salini S, Siletti E (2015) Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. ArXiv e-prints 1512.01569

Iacus SM, Porro G, Salini S, Siletti E (2017) How to exploit big data from social networks: a subjective well-being indicator via Twitter. In: Petrucci A, Verde R (eds) SIS 2017. Statistics and data science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Firenze University Press, Firenze, pp 537–542

Iacus SM, Porro G, Salini S, Siletti E (forthcoming) Controlling for selection bias in social media indicators through official statistics: A proposal. Journal of Official

Statistics

ISTAT (2017) La soddisfazione dei cittadini per le condizioni di vita. Tech. rep., ISTA, URL https://www.istat.it/it/files//2018/01/Soddisfazione-cittadini.pdf

Kahneman D, Krueger AB (2006) Developments in the measurement of subjective well-being. Journal of Economic Perspectives 20(1):3–24

Kwong BM, McPherson SM, Shibata JFA, Zee OT (2012) Facebook: Data mining the world's largest focus group. Graziadia Business Review 15:1–8

New Economics Foundation (2012) The Happy Planet Index: 2012 report. a global index of sustainable well-being. Tech. rep., New Economics Foundation

Rosembaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Schwarz N (1999) Self-reports: how the questions shape the answers. American psychologist 54(2):93–105

Schwarz N, Strack F (1999) Reports of subjective well-being: Judgmental processes and their methodological implications. Well-being: The foundations of hedonic psychology 7:61–84

Stiglitz J, Sen A, Fitoussi JP (2009) Report by the commission on the measurement of economic performance and social progress. Tech. rep., INSEE

## A - Appendix

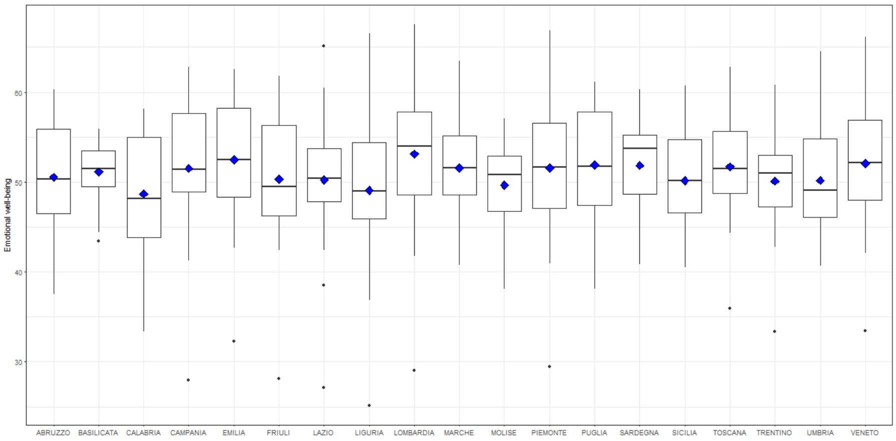

**Fig. A.1** Distribution of the emotional well-being (`emo`) for the Italian regions

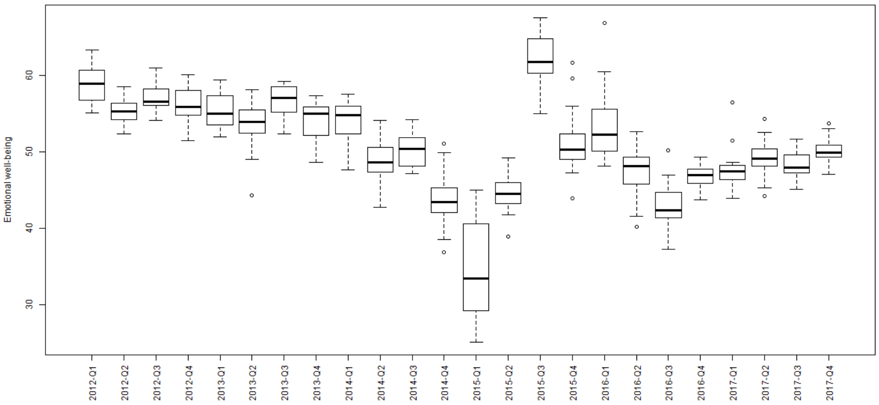

**Fig. A.2** Distribution of the emotional well-being (`emo`) in the 24 quarters from 2012 to 2017

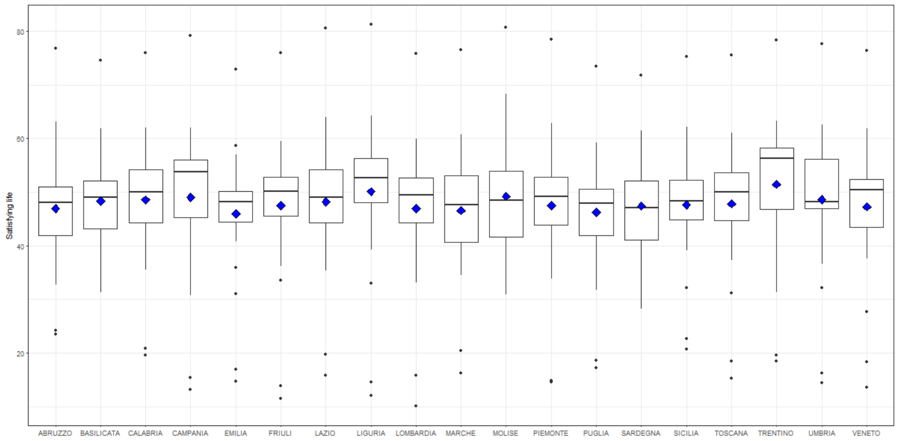

**Fig. A.3** Distribution of life satisfaction (`sat`) for the Italian regions

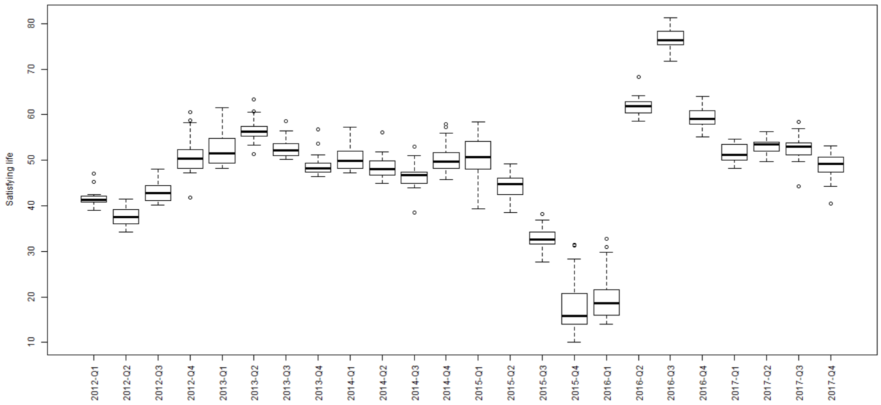

**Fig. A.4** Distribution of life satisfaction (`sat`) in the 24 quarters from 2012 to 2017

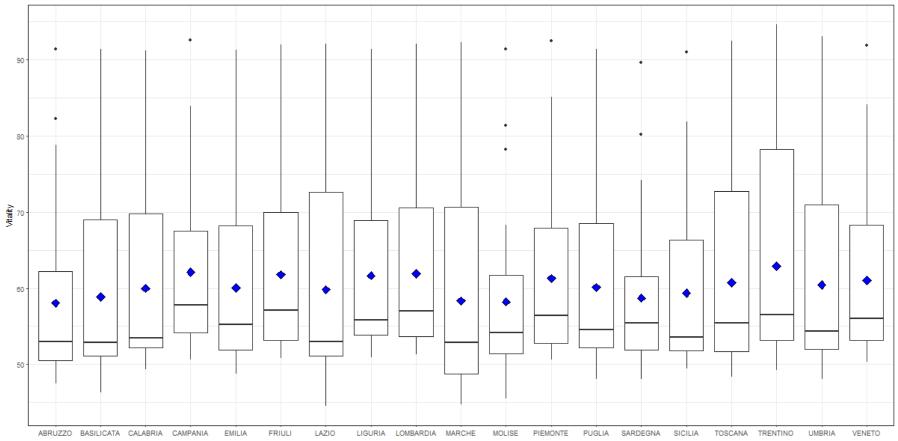

**Fig. A.5** Distribution of vitality (`vit`) for the Italian regions

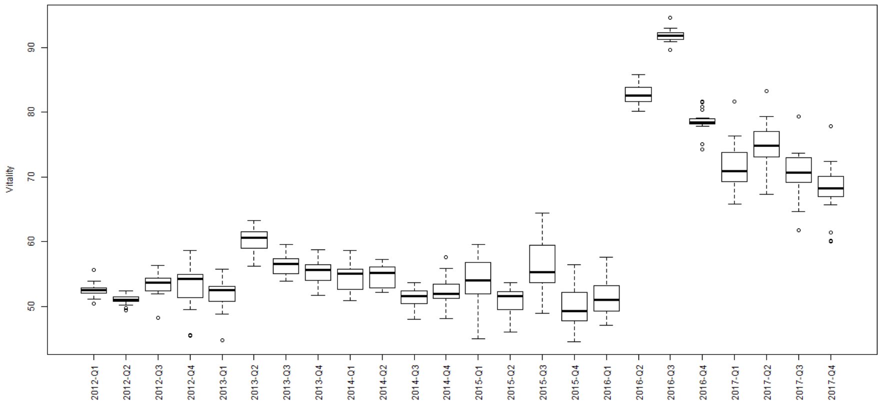

**Fig. A.6** Distribution of vitality (`vit`) in the 24 quarters from 2012 to 2017

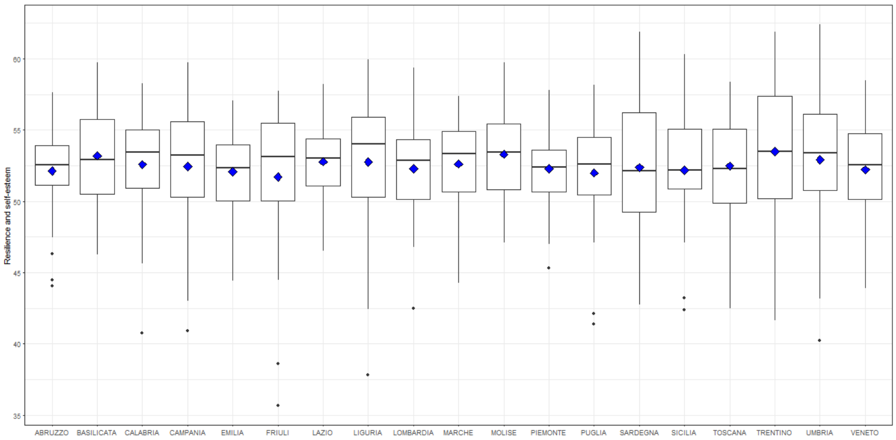

**Fig. A.7** Distribution of resilience and self-esteem (`res`) for the Italian regions

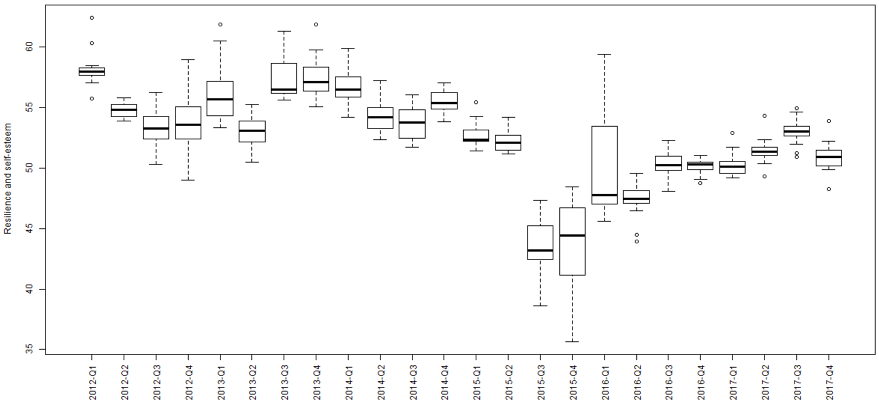

**Fig. A.8** Distribution of resilience and self-esteem (`res`) in the 24 quarters from 2012 to 2017

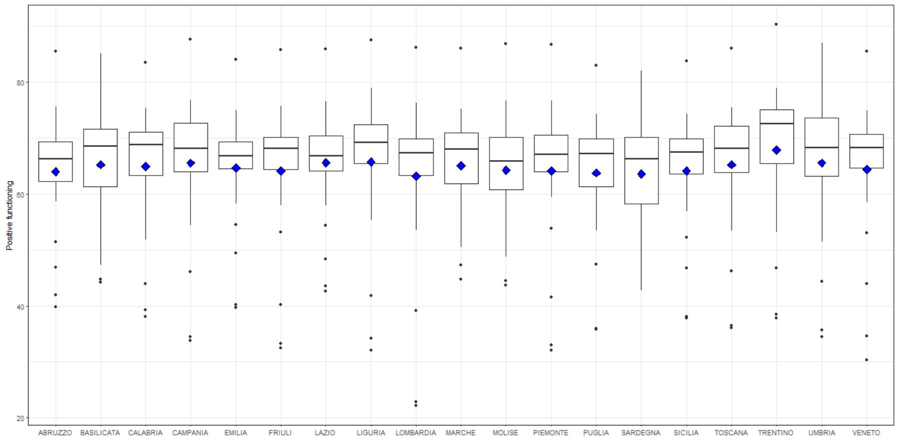

**Fig. A.9** Distribution of positive functioning (`fun`) for the Italian regions

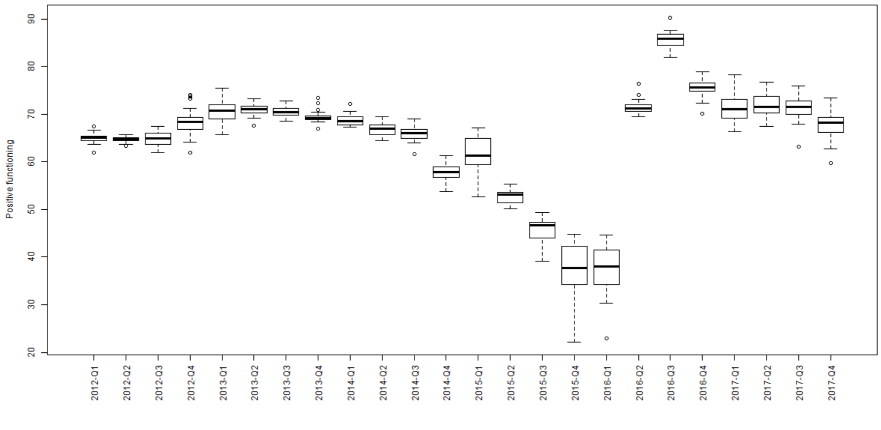

**Fig. A.10** Distribution of positive functioning (`fun`) in the 24 quarters from 2012 to 2017

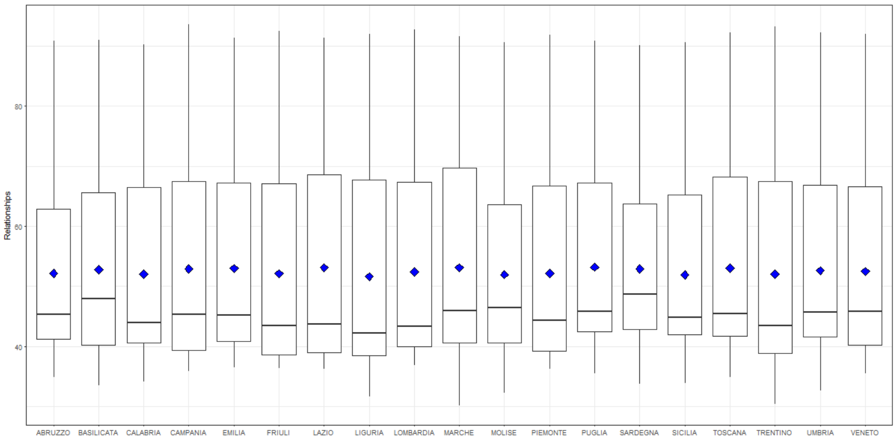

**Fig. A.11** Distribution of relationships (`rel`) for the Italian regions

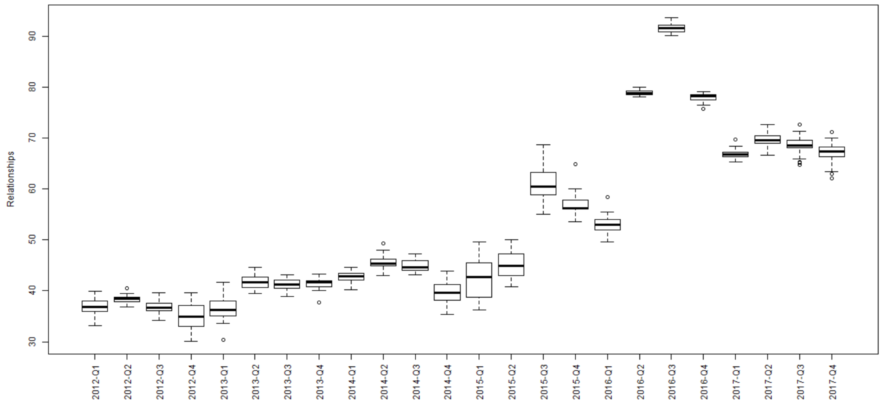

**Fig. A.12** Distribution of relationships (`rel`) in the 24 quarters from 2012 to 2017