Social Indicators Research
https://doi.org/10.1007/s11205-020-02581-8

ORIGINAL RESEARCH

Evaluating Rank-Coherence of Crowd Rating in Customer Satisfaction

Venera Tomaselli[1] · **Giulio Giacomo Cantone**[2]

Accepted: 27 November 2020

Abstract
Crowd rating is a continuous and public process of data gathering that allows the display of general quantitative opinions on a topic from online anonymous networks as they are crowds. Online platforms leveraged these technologies to improve predictive tasks in marketing. However, we argue for a different employment of crowd rating as a tool of public utility to support social contexts suffering to adverse selection, like tourism. This aim needs to deal with issues in both method of measurement and analysis of data, and with common biases associated to public disclosure of rating information. We propose an evaluative method to investigate fairness of common measures of rating procedures with the peculiar perspective of assessing linearity of the ranked outcomes. This is tested on a longitudinal observational case of 7 years of customer satisfaction ratings, for a total amount of 26.888 reviews. According to the results obtained from the sampled dataset, analysed with the proposed evaluative method, there is a trade-off between loss of (potentially) biased information on ratings and fairness of the resulting rankings. However, computing an ad hoc unbiased ranking case, the ranking outcome through the time-weighted measure is not significantly different from the ad hoc unbiased case.

Keywords Crowd rating · Ranking · Rank-coherence · Customer satisfaction · Tourism

1 Introduction: Rating from a Crowd

Crowdsourcing is generic terminology to categorise different practices in technological design and management. According to Estellés-Arolas and González-Ladrón-de-Guevara (2012), different definitions of crowdsourcing co-existed: some authors presented certain specific cases as paradigmatic, but no consensus was reached. The common factors are: (1) there is a multitude of individuals, (2) these individuals cooperate towards a common task

✉ Venera Tomaselli
venera.tomaselli@unict.it

Giulio Giacomo Cantone
giulio.cantone@phd.unict.it

[1] Department of Political and Social Sciences, University of Catania, 8, Vittorio Emanuele II, 95131 Catania, Italy

[2] Department of Physics and Astronomy, University of Catania, 64, S. Sofia, 95123 Catania, Italy

Published online: 05 December 2020

or a common goal, (3) these individuals are connected through a web technology ('platform') and generally they can mutually monitor each other (at least partially).

We propose to take into account the paradigm of Geiger et al. (2012). Authors proposed four "archetypes of crowdsourcing information systems": crowd rating, crowd creation, crowd processing and crowd solving (*ivi*, pp. 4–6).

In crowd rating the task is to bring "votes on given topics, […] such as a spectrum of opinions or collective assessments and predictions that reflect the 'wisdom of crowds'." (*ivi*, p. 5). Therefore, the crowd estimates a numerical value. Crowd rating's tasks are twice useful: (1) regard internal mechanics of a crowdsourcing information system because they allow the start-up of those advanced processes that demand quantitative parametrisation in order to properly work, e.g. recommender systems; (2) regard external value of those data, e.g. data can be sold or further researched, etc. These practices are commonly adopted by digital businesses and they are popular among researchers because the cost of deployment is lower than alternatives (Goodman and Paolacci 2017).

Dellarocas (2011) noticed at least two further sociological effects that contribute to crowd rating's popularity in digital economies: (1) an increase in trust into digital commerce through a sense of community; (2) a "lock-in" effect that makes less likely that a user involved in ratings will leave the platform for a competitor.

We noticed that according this definition of crowd rating the methodological focus is on the *crowd*. Crowds have two features: (1) are undefined, as no deep demographic information is usually collected from individuals; (2) are unfixed in the number, as the time in which the rating activities are performed is continuous and not constrained to a temporal limit, so new users can always join the crowd.

While some authors evaluate the impact of crowdsourcing a possible methodological revolution for any science involving problem-solving (Zheng et al. 2017), we think *customer satisfaction* is an excellent framework for development of methods based on crowd rating. Goodman and Paolacci (2017) report that there is a reasonable expectation that crowdsourced techniques will be "the routine" in customer satisfaction research.

The most noticeable propriety of the concept of *satisfaction* is that it can be interpreted as a finite and non-negative feature (Pizam et al. 2016). If this is the case, the dimension of satisfaction can be observed as a dichotomic variable [*satisfied* or *not-satisfied*] or as a continuous measure in the codomain [0,1], which represents the *degree of satisfaction* between no satisfaction and full satisfaction. The second is interesting in our opinion because the measures in [0,1] can be operationalised through normalised frequencies of past interactions, e.g. relative frequencies of ratings between customers and items (Proietti 2019). Then, in customer satisfaction's context this numerical value may be interpreted as estimated likelihood that the next customer will be satisfied through the item.

This interpretation of normalized values as predictors is paradigmatic in data science applications, as recommender systems or search engines. Recommender systems match data to the profile of a user, while search engines match data-structured answers to information science 'queries'. The value of these techniques is widely recognised but they require an adequate input of information and resources such as data storage and computational power (Varian 2016; Khusro et al. 2016; Melville and Sindhwani 2017).

Both recommender systems and web searches share a common feature: they need a procedure to order the ensemble of possible recommendations or answers in a ranking. While the general approach to both crowd rating and customer satisfaction tries to assert reliability of the rating measure as if the procedure is measuring a real (although often intersubjective) feature of the item, our intuition is to focus on internal coherence of rankings. In our approach ratings are only methodological tools without the necessity of being real indexes.

The present study is organized as follows: Sect. 2 covers the description of general concepts involved in this work such as the purposes of a rating system, the common assessments of customer satisfaction with data analysis, and theoretical context to develop a method of crowd rating; Sect. 3 presents and comments the development of the technical tools to reach results for crowd rating; Sect. 4 displays a real-world dataset scraped from an online platform; the Sect. 5 is dedicated to the analysis of the results over the sampled dataset; in Sect. 6 we provide final remarks on future developments.

2 Theoretical paradigm

We mentioned in Sect. 1 that data analysis is a core business process in web-based service platforms. Data analysis presupposes a method of data collection; therefore, the two operations share a common design.

Different platforms have different needs reflecting the proper nature of goods and services involved in their business. The platform Netflix is a web-based service which offers *streaming* of movies and shows in exchange of a periodic fee. Among many different business cases, there are good reasons to consider the case of Netflix's early system as paradigmatic for this field of research:

i. there is good literature coverage of the history of data analysis methods employed by Netflix because in 2006 the platform issued the Netflix Prize, a competition opened to anyone in order to improve predictive algorithms with a cash prize of $1,000,000. This event represents a paradigmatic historical record to understand the development of Big Data methodologies (Bell and Koren 2007; Koren and Bell 2015).
ii. features of movie ratings are often less problematic than alternatives *e.g.*, a *car sharing* service where local characteristics *e.g.*, quality of the roads, can affect the satisfaction of customer. Video streaming is also accessible from almost everywhere as it does not require a travel (differently from tourism).

In a platform like Netflix, data is managed in this way (Langville and Meyer 2012):

1. The platform has a list of *items.* Every *item* records the presence of an *atomic*, *standardized* good (eventually, a collection of goods, like *series*) that the platform is streaming. The platform controls the number of items at any time through the decision of what to put on offer and generally this number ends to be in the order of thousands of items.
2. The platform accepts the registration of *users* within. The *user* is the ID record of a personal *web account* of an individual. All the users form the list of users. The platform, under some circumstances, censors and remove a user but generally the platform cannot control the number of users, as it can increase at different rate of growth at any given time. Platforms has users in the order of millions.
3. Customers consume goods. Then, the platform asks the customer to declare their opinion on this experience i.e., *to rate*. This is not mandatory. The method to measure this opinion may differ in some features but the universal methodology adopts ordinal scales. The platform then proposes to the customer to explicit a value from a finite *multipoint* scale M *of scores*, which is a subset of ordered natural numbers, starting from 1 to *m*. This scale can be called 'numeric support'. Numeric support M: {*1, 2, …, m*} can be biunivocal associated to a semantic support, which links every numerical value (*scores*)

to a proposition in a natural language. Generally, M is expected to be *complete*, which means that it contains every natural number from 1 to m. The information system keeps track of when the value is submitted to the platform, therefore the dataset released to Netflix Prize was a vertical matrix with four variables (user, movie, score, date) and 100 480 507 observed cases.

4. The lists of items and users are converted into two vectors, U: $\{u_1, u_2, \ldots, u_i\}$ and P: $\{p_1, p_2, \ldots, p_j\}$ (p stands for 'product') and crossed in a in $\boldsymbol{X}$: UxP 'rating matrix' format. The numerical value from M that u indicate into p is then recorded as x_{ij}. If there is no value to associate, a *null* value ('missing value') is recorded instead. In a rating matrix, the data on time is not recorded.

For platforms, aim of data analysis is predictivity. There are two kind of inferences that can be made: (1) to predict a missing x_{ij} through an estimate before the customer rate the p item; (2) to propose an ordered list (*ranking*) of unexperienced *items* to a customer such that the top ranked items in the list have the highest likelihood to *satisfy* the taste of the customer. It can be noticed that (2) is the essential task of recommender systems. The focus is only on the accuracy of top ranked items, which are those displayed and *recommended* to the specific customer. This operation is called 'filtering' in this field of research.[1] Generally, the more accurate the (2) the better is for the platform. We noticed in Bell and Koren (2007) that the measures for (1) are actually proxies to determine (2), so we may call this *predictive ranking*.

Recommender systems are usually, but not exclusively, based on filtering algorithms commonly referred as *collaborative filtering* (CF). CF algorithms are complex and very technical, mostly because they are combinations of data analysis methods from different scientific fields such as Statistics, Biology, Engineering, etc. These methods are mostly employed with the aim of reducing the dimensions of *items* (Principal Components Analysis, Single Value Decomposition, etc.) and to infer information on *users* through data observed in similar users (Neighbourhood-based CF, *memory-based* algorithms). Once the estimates are reached for the X_i vector of the u_i user they are ordered in the ranking vector and displayed as recommendations (Koren and Bell 2015; Aggarwal 2016; Khusro et al. 2016; Melville and Sindhwani 2017).

Common issues in CF are:

- Sparsity of data: CF cannot well perform when there are too much missing values in the X matrix.
- Scalability: a CF algorithm can well perform with an amount of r and then much worse perform when this amount increases over time.
- Shilling: *robustness* against *frauds* is a desirable feature in a recommender system (Si and Li 2020).

From this description of the Netflix-paradigm we can extract three core assumptions. When these are satisfied, we expect a good performance from CF:

[1] This terminology is actually connected to the mathematical concept of dynamical estimation (Jazwinski 1970) but in this case it refers at the aim of the design of data analysis, which acts as a "filter" to provide best content to users.

– Customers do not change their opinions over time, or that at least they do not to this often. This implies that a change in x_{ij} should be treated as a material error and not a model error.
– Items do not change their substantial content over time. This is satisfied for movies: once a movie is published on streaming, rarely is *re-cut* or modified in any nontrivial way.
– The order in which the items are experienced by the same customer does not influence their internal process of rating. This means that stochastic processes of both rating items and being rated by users are supposed to be 'ergodic' (Sinai 1976). According to such propriety, given an enough number of ratings, the dynamic processes of giving a rate from own past performance reflect a real probability density function (PDF) of the variable 'satisfaction' of the item. Under this assumption, recommender systems work as if 'time is reversible'. If this is the case, we say that a time-estimator (in this case, 'performance') approximates an ensemble-parameter ('satisfaction').[2]

2.1 Satisfaction analyses in tourism: a public utility

We think that even after technical improvement of algorithms, assumption of time-reversibility in dynamical analysis poses supplementary issues for other business contexts. Tourism businesses can be very reactive to quality assessment and customer satisfaction. As a consequence, the system of rating can trigger a reaction into its subjects. If this is the case, predictions and recommendations are still factually feasible, but a descriptive assessment of the data seems a more appropriate approach.

We hold this opinion because 'hospitality business' have distinct and well researched features: tourism happens in geographical places which can be accessible or affordable only for some demographics. As a consequence, this makes 'localized sparsity' of data, which means that there are inequalities in how much information is collected about different subjects (e.g. restaurants, hotels, etc.). Some subjects are reliable and well-covered by public knowledge (often a minority) while others cannot reach minimum requirements to perform any data analysis (Lucas et al. 2013).

Incompleteness of information is also asymmetrically distributed between the seller and the (potential) consumer: the sellers know their goods in detail ("adverse selection", Akerlof 1970). In these asymmetric market situations, trust among economic agents is valuable. Consumers have found in services that provide public ratings (e.g. tourism guidebook) a tool to overcome this asymmetry in information (Clippinger 2011; Fernández-Barcala et al. 2010).

Kenett and Salini (2011) provide a compact review of established statistical approaches to customer satisfaction:

– Bayesian networks: this method builds a data structure called directed acyclic graph (DAG). DAGs are very useful for recommender systems because they simulate a decisional circuit. This approach is accurate in investigation of causes and effects behind data observed in customer behaviour. It is a predictive approach with a focus on social mechanics.

[2] Even within the context of streaming business, developments in CF algorithms tried to overcome the assumption of 'time-reversibility' in order to improve predictions (Koren 2010; Koren and Bell 2015).

– Generalized Mixture Models (GMM): this is the name for the family of Mixture Models, or probability models that combinate different distributions. Authors reference Combination of discrete Uniform and Binomial (CUB), originally proposed by Piccolo and D'Elia (2008) and Iannario and Piccolo (2010), as the model suited for customer satisfaction. CUBs are modeled after general psychological arguments on question 'how the raters are going to process their choice of score?'. CUBs estimate two parameters of the rating process, ranged in [0,1]: ξ for *feeling* and π for *uncertainty*. As we understand this model through the insights for longitudinal analysis of Proietti (2019), we think it is very flexible: ξ can be estimated through a vector of assigned values from a single user to a subset of ensemble P, but also through a vector of values referred to a single item by a subset of users.
– Rasch models (RM): De Battisti, Nicolini, and Salini (2010) proposed their adaptation of RM for customer satisfaction. Their model estimates *satisfaction* of a user and *quality* of a good through a rating matrix where users are the rows and questions about the goods are the columns. In our opinion, the original RM is among the first attempts to formalise an analysis on a format which shares similarities with the **X**: UxP rating matrix of recommender systems. But the premise to adoption a RM is to collect a not *sparse* matrix.

CUB seems the more promising *theory* in the peculiar context of crowd rating because it holds less assumptions on the availability and format of data. This method balances its actual *rating* parameter (ξ) through the output of a second value of uncertainty in π. CUBs can be a useful tool to describe a population or a sample of items through a Cartesian chart of feeling and uncertainty. Proietti (2019) and Piccolo and Simone (2019) argued on dynamical proprieties of CUBs parameters when applied in dynamical analyses, which highlighted a noteworthy propriety of ξ:

$$\pi = 1 \rightarrow \left(1 - \hat{\xi}\right) = \frac{\bar{X}_p - 1}{m - 1} \tag{1}$$

where the right hand of the equation is the mean of values after the normalisation that we mentioned in Sect. 1, and the estimator of ξ is the maximum likelihood estimator.[3]

Moreover, all the three approaches are founded on the assumption that a *good* (i.e. an *item* in crowd rating) can be observed through multiple questions and that sociodemographic data on users are available: "In such surveys, customers are requested to fill in questionnaires with typically 10 to 80 or even 100 questions" (Kenett and Salini, p. 465). This condition is often not satisfied in crowd rating.

Dynamical modelling of CUBs opens for the possibility to accept a single measured value (i.e., a single question) from a multipoint scale as an accurate estimate of a 'satisfaction' variable. However, according to Pizam et al. (2016) this seems not the case. Even the American Customer Satisfaction Index adopts 3 dimensions of performance in order to reconstruct satisfaction as a latent variable (Fornell et al. 1996). We think that the difference between 'overall' satisfaction, which is the latent construct of the *sentiment* about the past experience, and 'future behavioural intention' or 'feeling towards recommendation'

[3] In the original paper of Piccolo and Simone (2019) the left hand of the equation was only the estimator of ξ, while Proietti referenced correctly to its complement. This happens because, in the Proietti's dataset, numeric values had an inverted semantic support (e.g., 1 was *better* than 2). In absence of semantic interpretation, $1 - \xi$ is actually the measure of *feeling*.

(e.g., the answer to the question: "would you suggest this experience to a friend?") is relevant but only for predictive analysis.

Pizam et al. (2016) conclude their work recalling the necessity of a transition to web-based methods of data collection. We notice that the general format of web-based rating systems is often made of few variables if not only one variable. This seems a radicalisation of what Robert Groves (2011) argues being a historical trend of social research. This trend was studied by Goodman and Paolacci (2017), too. However, this topic seems not conclusive from a theoretical standpoint: multiple questions are still possible in a crowd rating information system and Leal et al. (2018) argued that multivariate approaches improves predictivity of recommender systems.

2.2 Crowd rating for customer satisfaction in tourism: biases and methodology

There are evidences that under the assumption of mutual independence of judgements, estimations made from the opinions of a crowd of N judges can be accurate (Galton 1907; Wallis 2014). This methodology always attracted the interest of statisticians but experimental and quasi-experimental research over this topic surged after 2000 (Ariely et al. 2000; Soll and Larrick 2009; Müller-Trede et al. 2018). However, we have to highlight major differences in core assumptions between experimental methodology and crowd rating information systems (observational methodology):

- Crowd rating lacks metric values: Galton asked people to estimate a *physical weight* through an unconstrained metric, which is the reason why he employed the median and not the mean, we guess. Crowd rating aims to estimate features like *quality* or *satisfaction*. The established method to evaluate an intersubjective value is the use of ordinal multipoint scales. These scales are commonly employed in online rating systems.
- In crowd rating there is no secrecy of opinions: an experiment is structured to have a start and an end, and generally intermediary results are kept in secrecy to ensure control over biases. While open platforms vigorously enforce secrecy on how their algorithms are coded, their business models are still based on showing to public data including comments, reviews, and ratings.
- A crowd rating information system enables a competition: it is common knowledge in digital marketing that when a new technology enables to rank products under a common criterion of database interrogation ('query'), business aim to keep a top ranking for their ranked products and brands (e.g. Search Engine Optimization practices; Varian 2016).

These features seem to enact those sociological effects already mentioned in Sect. 1 that influence people's behaviour within crowd rating information systems (Jeacle and Carter 2011; Érdi 2019). Unfortunately, biases are commonly observed in quantitative evaluation of satisfaction by crowds, too:

i. Non-independence of observations: we have evidences that *public* rating systems are not ergodic. Both experimental (Salganik et al. 2006) and observational (Lee et al. 2015) studies on crowd rating information systems suggest that, in the absence of secrecy of 'what is trending', judgements over products converge or "do herding" (*ibidem*) towards a strong modal class of answers. Further research on platforms Amazon and Yelp confirmed the hypothesis of the existence of a social mechanism

of herding ensuring that earlier ratings are more likely to influence future 'popularity' of products than later ones (Bai et al. 2018).

ii. Survivorship bias: competition among what is subjected to being rated reflects competition for survival in a market (Farmer 2011). Through this struggle for survival, some subjects may disappear from the market, and others may show up. Not only subjects in the same *query* or ranking have different lifespans, but also their data can be retroactively censored by platforms, for the reason that the platforms do not desire to host an inactive or misleading subject in their rating system. This could be impactful both in predictive and descriptive analysis because it censors those subjects where it is more likely that 'unpopularity' and *low satisfaction* will be observed. More generally, this statistical feature takes name of *survivorship bias* and it skews the distribution of ratings into higher numerical values (Mangel and Samaniego 1984).

iii. Frauds and optimization strategies: sometimes platforms lack clear procedures to confirm the general *sincerity* of the submitted data. While technologies to improve *fake detection* are constantly in development, frauds (*shills* in technical jargon; Si and Li 2020) are usually a consistent factor of skewness in reviews (Ott et al. 2012; Li et al. 2014). A further reflection is necessary: while a subject who actually manipulates a ranking through the submission of *fakes* may be held responsible of crime, TripAdvisor states that 'optimization' and anything that does not involve a 'payment' to fake a review is not against its Terms of Service.[4] We could conclude that 'asking gently' to submit a max-scored rating should be considered a legitimate strategy of optimization of reputation and awareness, thus introducing another potential bias.

Differences between experimental and observational methods in social science are largely discussed, so that Salganik (2018) remarked the importance of this issue in the "Digital Age" of social research, with the social experiments focused on social mechanics and predictions and observational studies focused in population inference and description of common, public features. Given the nature of open data crowdsourced through an information system, we focus on descriptive proprieties of the ranking systems. In our opinion rankings are the most important structure of a public crowd rating information system because their benchmarking function is the main factors of aforementioned biases (Érdi 2019; Mari and Ruffini 2018).

Brief considerations leading our proposed method of descriptive analysis for rankings:

a. Locality: tourism happens in geographical places; businesses located in distant touristic localities are often not comparable.
b. Publicity: recommender systems collect information from a crowd and then usually propose individual suggestions. But we came to conclusion that tourism economies benefit more if information systems display *public* information and not personalised information. Hence, we have in mind a system that offers suggestions about queries on well constrained geographic area.
c. Descriptive use: recommendations happen in the practical format of rankings, which are associations between natural numbers and a list of items. Rankings have the purpose to describe evidences of the past. Not all ranks have the same importance. Usually top ranked items are more important than others. By this we do not mean that the top items

[4] https://www.tripadvisor.com/TripAdvisorInsights/w3703.

associated to top ranks are strictly better but also that observers have deeper psychological commitment to know if these top items are actually 'good' than if average items are 'truly average'. We suspect, due to the psychological limit of memory and attention of *availability bias* (Tversky and Kahneman 1974), that the higher the number of ranks in the query, the more important are top ranked. This enacts the competition among businesses (ii.) to reach top ranks in public displays, e.g. queries in the form "best restaurants in this area". Rankings are not predictions, although inferential power of rankings and ratings can be tested and asserted (Alvo and Yu 2014; Corain et al. 2016).

d. Proprieties of measurement: rankings need a measurement procedure (Krantz et al. 1971) that works as the ordering criterion. In the contexts of a public ranking of tourism this measure is a *rating statistic*. One may be led to think that a rating has a real interpretation but in this context the rating is only a methodological tool. When a rating is the numerical output of a statistical procedure that takes as input many values provided by a crowd, we call this whole process a proper *crowd rating*. There can be different crowd rating statistics and some of them fit better the linear scheme of rankings.

e. Rank-coherence: rankings are updated over the time on a continuous basis. The permutation of ranks between two rankings generated through the same query but in two different times should describe properly what the system observed over the passage of time and not be influenced by the rating procedure (Corain et al. 2016). In other terms, when the task is to measure a change over the time, a system can burden only a limited degree of change of its method of measurement if it aims to be a reliable resource for the public. We define *strong rank-coherence* (i.e., linear homomophism) the condition supporting the proposition: for each possible couple of ranks, their distance divided through the distance of their associated ratings is constant (Krantz et al. 1971). This is an ideal condition to easily interpret the meaning of a ranking: strong rank-coherence or *linearity* implies that knowing both the rank and the rating of just an item from a query one can knows all the information on the rating scheme in the query. We refer to the condition of a rating system trending over the time into a strong rank-coherence as *weak* rank-coherence.

f. Fairness: weak rank-coherence implies that knowledge on the public value of the items comes with a degree of reliability. In other terms, a reliable ranking system must be *fair* about how it orders the items, even if the ratings can be not accurate about what they are measuring. This is because we do not assume rating values as real values but only as a tool of measurement. An interesting propriety of this definition is that *ex aequo* are the worst possible outcomes of a ranking in terms of fairness.

3 Methods

In Sect. 2 we noticed that Netflix released their data collection in a form of a matrix, or 'dataframe' made of 4 columns, each for a variable: user, item, score and time. Even if in more complex data structures all this information (and probably even more!) is employed, in the basic rating matrix **X:** UxP the information about time is lost, because time is assumed to be reversible. The ranking model of the recommender system is time-independent, then.

In our method time is not reversible but users are assumed mutually fungible. Hence, we say that we take-in time, take-out individuality of raters. The basic data structure is then a simple dataframe with three variables: time, item and the value (i.e., the *score*) of the

single instance rating from the multipoint scale of scores. Time is a variable that must be coded according to a 'time-step': an interval of time that works as a bin for all the recorded cases happening within its range. The standard time-step is the day, but alternatives can be weeks, months, quarters (e.g., Q1 = from 1st January to 31st March) or years. Weeks make an exceptions in how are referenced because while is usually convenient to reference the time-step with a calendar (e.g., t_{month} = 'January 2020'), weeks are conveniently referenced in relationship to the procession of time in the dataset (e.g., t_{week} = 100th).

We will refer to the generic time-step as t, the generic item as p, and the assumed valued as x. We will refer to the set of elements of the list of items as *ensemble* P and to the set of values associated to a generic p as a vector X_p. The multipoint scale is still referred as M, and its maximum value is m.

Once the time-step is chosen, the rating matrix **R:** TxP takes this format:

Where $r(p_{t,j})$ is a numerical value assumed at time t associated to item p_j through a function r ranged in [0,1]. We call $r(p)$ the *rating statistic* of X_p. This format reflects the practical need of anonymity of rating process. The rating function orders the P ensemble through $\{1, \ldots, \text{card}(P)\}$, from the highest to the lowest $r(p)$. The natural number associated to p is its rank, recalled through function rank(p).

3.1 Rating Functions

The debate about the estimator of central value of an ordinal variable is an open controversy. Velleman and Wilkinson (1993) reconstructed the start of the debate in 1953 with a written informal confutation of the notorious theory of *scales of measurement* (Stevens 1946) from the statistician Frederic M. Lord. According to authors, after the formalization provided by Luce (1959) and Krantz et al. (1971) the scientific consensus seemed to lay in favour of Steven's suggestion to not adopt arithmetic mean as estimator of central value for ordinal data. But in their opinion, this consensus was correctly questioned with further arguments provided by Rasch (1961), Lord and Novick (1968) and then after a decade by Guttman (1977) and Mosteller and Tukey (1977). After the 1990, the theoretical debate was still open, but practitioners rather often chose the mean over the median. This trend was particularly relevant in the rising digital industry (Lewis 1993; Lewis and Sauro 2016, pp. 250–254). To put things in perspective, even the American Customer Satisfaction Index is actually an average (Fornell et al. 1996).

Our choice goes into not adopting median. Median is an estimator with the propriety to be robust to skewness towards extreme values when data are unconstrained (e.g., as in Galton's experiment, see 1. in Sect. 2.2) but in our method data are constrained. On the other hand, median's lack of sensitivity towards small differences of value is a disadvantage in those cases where these differences, even the smallest, are decisive: a minimal increase of the median of X_p can cause a more than proportional difference in the form of a permutation in the ranking. This property is undesirable because it goes against the criterion of rank-coherence (see, e. in Sect. 2.2).

To give a perspective, in Baccianella et al. (2009) the sum of amounts (n) of $x = 4$ and $x = 5$ always had a frequency over to 0.7 of X. Hence, to rank items through the median always would have produced a binary classification in their ensemble (i.e., median was equal to 4 or equal to 5). When all of top ranked items are rated as $\text{Med}(X_p) = m$, they all score $\text{rank}(p) = 1$ *ex aequo*. If this is the case, they are all tallied in the first *bin* and SRLE reaches its maximal value (see, Sect. 3): the ranking format is not descriptive of the collected data. This does not mean that two or many items should not have the same rating,

but that a ranking is not a well-suited format for ensembles with too many *ex aequo* among top ranks.

We then propose the normalised mean in [0,1] as rating function. We will refer to this measure as 'Overall Satisfaction Average' (OSA):

$$\mathrm{OSA} = \frac{\bar{X}_p - 1}{m - 1} \tag{2}$$

The [0,1] constrain asks for a further interpretation of OSA. As we mentioned in Sect. 1, this constrain is characteristic of both relative frequency and likelihood. In our opinion a correct interpretation of OSA is that this measures the fraction of observed satisfaction scaled to its maximum value.

According to Langville and Meyer (2012, pp. 147–149) a rating statistic can be weighted to t ("weighting scheme"). A distinction is between continuous and discrete ("step-weighting" according to authors) schemes of weight. A continuous weight is a continuous function in t, which can commonly be linear, logarithmic, exponential, or logistic. Step-weighting chooses a normative interval of time $\Delta(t)$ meaning 'Δ time-steps before t'. So, the w weight of x observed before $t - \Delta$ is equal to 0, while if x is observed in the time between t and $t - \Delta$, w it is equal to 1. Δ is a normative interval in the sense that is not fully arbitrary but it reflects a practical consideration about the operating context e.g., if t is expressed in days, Δ could be 7 days (days in a week) or 91 days (the equivalent in days in a quarter). We think that a well-thought normative approach requires less statistical assumptions and is less prone to errors in modeling the continuous scheme. When OSA is step-Weighted (WOSA) the measure is the following:

$$\mathrm{WOSA} = \frac{\frac{\sum w_t x}{N_{\Delta(t)}} - 1}{m - 1}; \quad w_t = \begin{cases} 1; & t - t(x) < \Delta(t) \\ 0; & t - t(x) \geq \Delta(t) \end{cases} \tag{3}$$

We think 91 days as Δ is the most attuned normative solution for tourism. It can be interpreted as a 'shift of seasons' i.e. at the last day of a quarter or season this weight would take into account only data happened in the last quarter or season.

It can be noticed that only a user submitting $X_p = m$ shows full satisfaction from the p item, thus the rate of Full Satisfied (FS) users is the proportion:

$$FS = \frac{n|X = m}{N} \tag{4}$$

3.2 Evaluative Method

Proprieties of a ranking can be studied through pairing of the two functions that characterize it: $r(p)$ and rank(p). A noteworthy propriety of the rankings is the linear coherence of ranks (see, e. in Sect. 2.2): strong coherence of ranks is reached in a ranking when $r(p)$ is a linear transformation of $\mathrm{rank}^{-1}(p)$.

We want to evaluate top ranked items because they are the most important (see, c. in Sect. 2.2). To evaluate the coherence of top ranked items we adopt the measure Root Sum squared linear Ranking Error (RSRE) in k integer.

In order to compute this measure:

1. k is chosen as the last rank to be considered a 'top' rank;
2. the range $r(p_{\text{rank}=1}) - r(p_{\text{rank}=k})$ is split in $k/2$ intervals of equal length q;
3. The range $(r(p_{\text{rank}=1}) + q/2) - [r(p_{\text{rank}=k}) - q/2]$ is split of $[(k/2) - 1]$ intervals of equal length q;
4. each one from the $k - 1$ union of intervals is a *bin*. Each bin, aside first and last, is the intersection of 2 other bins. Under assumption of local linearity in the k top ranks, each *bin* tallies 2 ranks.

The RSRE is the root of the sum of the squares of the deviations in observed values in the bins:

$$\text{RSRE}_k = \sqrt{\sum_{\text{bin}=1}^{\text{bin}=k-1} [\text{card}(\text{bin}) - 2]^2} \tag{5}$$

There are k values of rating and $k - 1$ *bins*. RSRE_k reaches its maximum when $(k–1)$ rating values of r are tallied into a bin and 1 rating value is tallied into another:

$$\text{Max}\left(\text{RSRE}_k\right) = \sqrt{(k-3)^2 + (k-3) * 4 + 1} \tag{6}$$

where $(k - 3)$ is the deviation of the bin which tallies $(k - 1)$ rating values, 1 is the squared deviation of the bin which tallies 1 rating value, $(k - 3)$ is the amount of bins which tallies 0 ratings, and 4 is the square of the $(0 - 2)$ deviation. Therefore, the τ-statistic measures the rate of RSRE on its theoretical maximum value in k:

$$\tau = \sqrt{\frac{\sum \left[\text{card}\left(\text{bin}_{1\to(k-1)}\right) - 2\right]^2}{(k-3)^2 + 4(k-3) + 1}} \tag{7}$$

If one holds assumption of *flat* PDF of the model of perception of the M scale, τ is valid both to estimate trend in rank-coherence of k top ranked items and of P whole ensemble. If for the same ensemble of k-top items the associated value of τ over the time for a rating statistic r_1 are significantly higher than those for rating statistic r_2, then r_1 is less rank-coherent than r_2. We know that in crowd rating for $t \to \infty$ both $\text{card}(P) \to \infty$ and $\text{N} \to \infty$ (see, Sect. 2), therefore a trending decrease of τ over time shows that a rating statistic is rank-coherent for the purposes of crowd rating (see, Sect. 2.2).

But if one assumes that the PDF of the perception of M is Gaussian-shaped (i.e., nominal ogival), then no ranking is immediately a fair descriptive format of the changes in the P ensemble. However, for $\text{card(P)} \to \infty$ the model holds its validity due to the flatness of Gaussian's extreme tails after standardisation. In simple terms, this methodology is well suited for large ensembles (e.g. 1000 or more items).

Our assumptions on the PDF of perception of the scale: (1) is symmetrical; (2) is ogival shaped (i.e., not convex, not parabolic, etc.) or flat; (3) is platykurtic. In a crowd rating information system values of X are constrained and we assumed *completeness* of the scale M (see, Sect. 1). We have no explicit reasons to hold the assumption that customers will perceive this scale as not symmetric, although biases could be for reasons mentioned in Sect. 2.2. In other terms SRLE measures how well the rating function produces coherent rankings even in presence of not psychological biases in the dataset. We think that the experiment in Salganik et al. (2006) supports the assumption that is not the individual *perception* of the measuring tool the source of bias in the measure but other social mechanics.

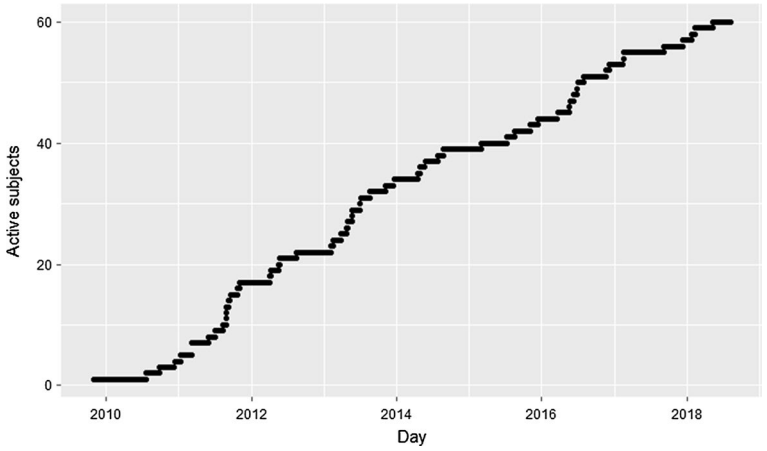

Fig. 1 Number of active subjects ('items') by day

4 Web-scraped Dataset

In Sect. 2.1 we reached the conclusion that crowd rating is suited to provide a ranking of a list of business from a query in a tourism location. According to this rationale, businesses should be comparable economic competitors on the same area. In the web platform TripAdvisor, users can submit a review of a restaurant. To each review, a numerical score from 1 through 5 is associated. We selected a tourism city as area and on August 5th 2018 we automatically collected data from TripAdvisor through an scraping algorithm in software language R (*rvest* package). This algorithm extract records from the list of the sampled web addresses (URL). This list was a sample of 60 URLs, from this query:

– Addressed in tourism city of Catania, IT.
– A restaurant with 'pizza' in the menu.
– had at least 20 reviews previous on August 5th, 2018.

The third condition acts as a filter to avoid unsubstantial *items*. Without the condition of 'at least 20 reviews' the sample would have been a hypothetical population of 225, that means that 165 restaurants with pizza in menu had less than 20 reviews. The first review in the sample happened on October 10th, 2009: 3204 days of activity. Restaurants with at least one rating at a time-step (day, week, etc.) is referred as an 'active subject' from that time-step, while restaurants with no reviews are considered 'inactive subjects' at that time-step. All cases of review ($N = 26.888$) were prompted into the dataframe (see, Sect. 3): Time, Item (restaurants) and Score (values of X). We estimated that N is more than at least 90% of whole reviews within the hypothetical population.

While the number of active subjects grows linearly (Fig. 1, R^2 of linear model with intercept = 0 is equal to 0.79), the number of collected reviews was not fully linear (Fig. 2, R^2 of linear model with intercept = 0 is equal to 0.51).

The maximum divergence between the two growth ratios of active subjects and collected reviews is reached on December 18th, 2013 (1513rd of 3204 days, 47% of time), when 34 of 60 subjects (57%) were already active.

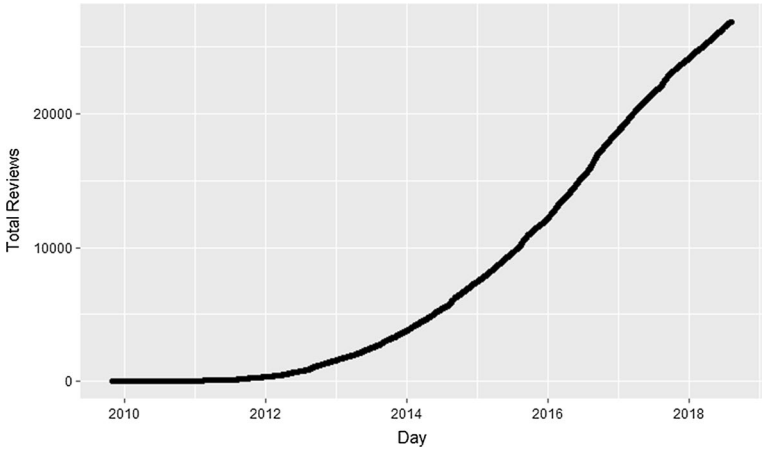

Fig. 2 Number of collected reviews by day

Table 1 scheme of rating matrix **R:** TxP

Time	P_1	P_2	…	P_j
$t=0$	$r(p_{0,1})$	Null	…	
$t=1$	$r(p_{1,1})$	$r(p_{1,2})$	…	
$t=2$	$r(p_{2,1})$	$r(p_{2,2})$	…	
…	…	…	…	…
t	$r(p_{t,1})$	$r(p_{t,2})$	…	$r(p_{t,j})$

If t is setup in days as time-step, on the whole dataset time-cumulative frequencies (i.e., adding the sum of absolute frequency of all previous time-steps to frequency of t, from $t=0$ to t, see Table 1 in Sect. 3) of the five classes of scores were stable from mid-August 2011 (Fig. 3 and Table 2).

August 7th, 2011 happens after 647 (.202) of 3204 days of activity, however only 0.005 of N was collected at that date (see, Fig. 2). This day was the first Sunday of August 2011 and August 5th, 2018 was the first Sunday of August 2018. So almost exactly 7 years passed between these two dates.

The modal class of X was $x=5$ since August 7th, 2011, floating around a median frequency of 0.441 (Table 2). Therefore, we suspect median of scores can be a not good statistic for ranking (see, Sect. 3.1). Data are consistent with results from previous studies on Italian cities on TripAdvisor (Baccianella et al. 2009). We questioned if this feature was independent from the time-step format. Switching to a week format by *binning* all the reviews from a Sunday through its subsequent Saturday, starting from August 7th, 2011 and ending August 4th, 2018 (7 years, 365 weeks), we got a confirm about the hypothesis of independence from time-step: in the weekly format, $x=5$ was still modal and had a time-median of f(x) equal to 0.449.

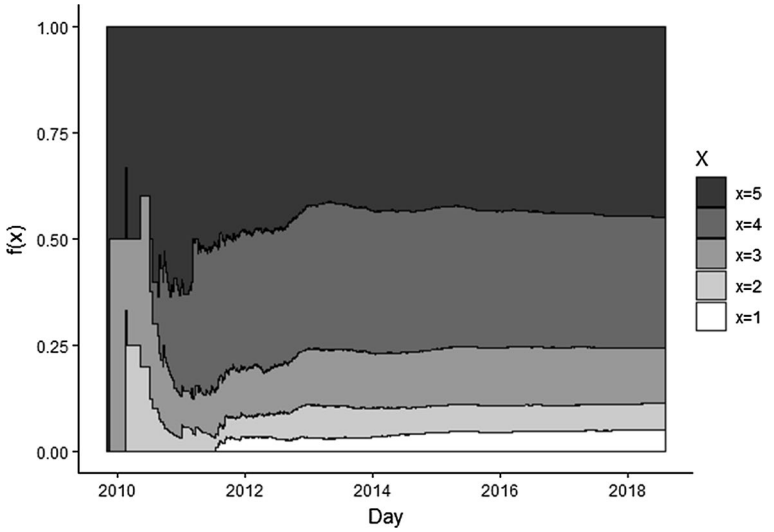

Fig. 3 Time-cumulative frequencies of scores

Table 2 frequencies of the class of scores on August 5th 2018 (N = 26.888), and time-median over the time

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$
$n(x)$	1395	1654	3509	8190	12 140
$f(x)$	0.0519	0.0615	0.130	0.305	0.451
Time-median of $f(x)$	0.0362	0.0620	0.133	0.320	0.441

5 Results

Section 3.1 presents three rating statistics (OSA, WOSA, and FS) and Sect. 3.2 presents the proposed τ-statistic in order to evaluate rank-coherence over time of ratings statistics for a queried dataset. Before to provide results of the evaluation of aforementioned statistics, we provide two arguments on the reason we think evaluating performance of OSA is important:

i. arithmetical mean is widely discussed statistic in theoretical studies. In Sects. 2.1 and 3.1 is noticed that there is a structural relationship between CUB methodology and OSA: *feeling* parameter ξ tends asymptotically to (1 – OSA) under some assumptions of PDF.
ii. the average often has pivotal role in practical applications of rating systems (see, Sect. 3.1). This seems true for the sampled dataset (see, Sect. 4), too. We collected a ranking (Ranking A) of 60 items on August 5th, 2018. Ranking A is the observed ranking on the platform TripAdvisor from the query, sorted through the recommender system (see, Sect. 2) of the platform. Rating B is the ranking from sorting the same 60 items applying OSA as rating function, instead. We found that the Spearman ordinal correlation between Ranking A and Ranking B was 0.54. But the Spearman

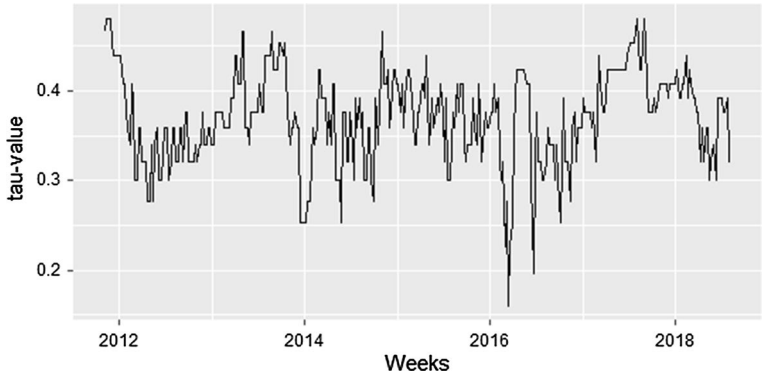

Fig. 4 Values of τ of OSA, week by week over 7 years of activity

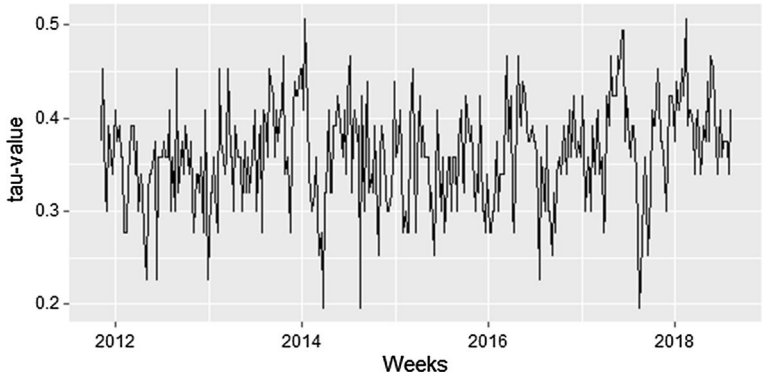

Fig. 5 Values of τ of WOSA, week by week over 7 years of activity

correlation raises to 0.97 when pairing only those items which scored a higher rank in B than in A (higher rank means *worse* performance), and decreases to 0.51 for the items which scored a lower rank in B than in A. This leads to think that the ordering criterion in the recommender system of TripAdvisor is not dissimilar from an averaging function of crowd's opinions but at the same time the procedure it is adjusted through an eventual smoothing down of outliers with a higher value of this rating.

Therefore, as a general rule, OSA is a good benchmark to evaluate alternative rating functions.

We computed $\tau_{k=10}$ of OSA, WOSA, and FS for each week from Monday 7th November, 2011 through Sunday 8th, August 2018, for a total of $t=353$ weeks (for OSA, see Fig. 4; for WOSA, see Fig. 5; for FS see Fig. 6). This t happens 12 weeks (=91 days) after August 7th, 2011, so it is attuned with the w of WOSA. The choice of $k=10$ is arbitrary, however it reflects a realistic range of comparison between items before falling outside psychological *availability* of benchmarking alternatives (see, c. in Sect. 2.2).

If a time series has a decreasing τ value over time, we accept the rating statistic as a rank-coherent procedure to sort a queried list in a ranking (see, Sects. 2.2 and 3.2). It is evident from Figs. 4, 5, and 6 that the *noise* component of the time series is dominant over

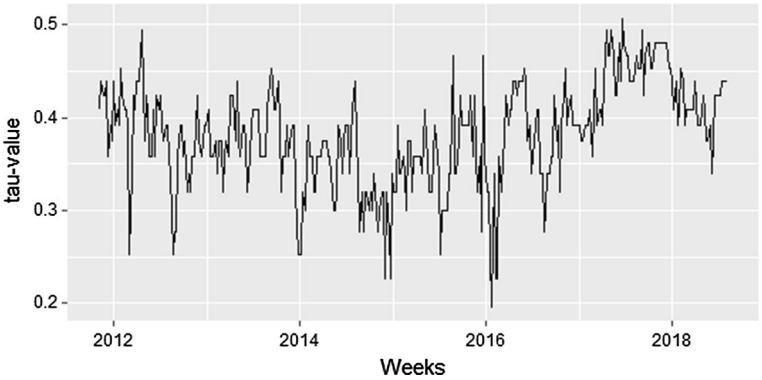

Fig. 6 Values of τ of FS, week by week over 7 years of activity

Table 3 Statistical tests for distributions of time series of τ

Null Hypothesis (H_0)	Median test (Wilcoxon) *p*-value	(H_0)	Kolmogorov–Smirnov test *p*-value
$Med(\tau_{WOSA}) - Med(\tau_{OSA}) > 0$	.0052 (Reject H_0)	$\tau_{OSA} > \tau_{WOSA}$	.9557 (Do not reject H_0)
$Med(\tau_{FS}) - Med(\tau_{OSA}) > 0$	.9981(Do not reject H_0)	$\tau_{OSA} > \tau_{FS}$	.0067 (Reject H_0)
$Med(\tau_{WOSA}) - Med(\tau_{FS}) > 0$	.0000 (Reject H_0)	$\tau_{FS} > \tau_{WOSA}$	.9972 (Do not reject H_0)

a trend, therefore the result is that neither OSA, WOSA, nor FS are fair rating statistics for the query.

If a time series has lower values of τ than another, the rating statistic is better performing with the aim of reduction of rank-incoherence over time of rating values and therefore is fairer.

The statistical interpretation of Table 3 is that $\tau_{FS} > \tau_{OSA} > \tau_{WOSA}$ is the only consistent sorting of the rating statistics from dataset's evidences. Is this sorting due to intrinsic proprieties of the measures, or to data? In Sect. 2.2 we highlighted major factors of expected biases: (1) optimization towards *m* class of score (e.g. asking "would you gently rate 5 if you liked the meal?"; which is not *shilling*, i.d. fraud methods); (2) competition, skewing results towards higher values of *x*; (3) *herding*, which reinforces effects of (1) and (2). The logical conclusion from this set of biases is that *m* will be the most biased class of scores. This is reflected in the sampled dataset, for $m = 5$ (see, Sect. 4).

In order to test if expected biases can affect rank-coherence of rating statistics we computed an Ad Hoc (AH) statistic:

$$AH = \frac{n|x = 4}{N - (n|x = 5)} \tag{8}$$

which is a rating measure purely functional to test effect of the expected biased class in the data: $x = 5$.

AH is the same format of rate n(m)/N of FS as if, hypothetically, all the information on $x = 5$ would be lost from the records. Previous tests indicated that for these data $\tau_{FS} > \tau_{WOSA}$. However, having observed the values of τ_{AH} (Fig. 7), a Kolmogorov–Smirnov test does

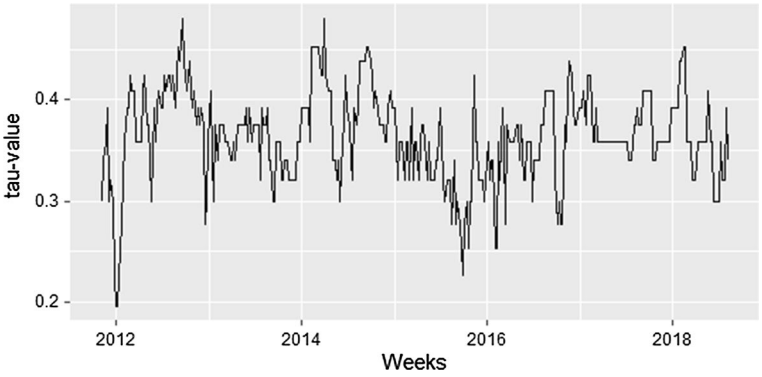

Fig. 7 Values of τ of WOSA, week by week over 7 years of activity

not reject the hypothesis that there are no significant differences between the cumulative distribution functions (CDF) of τ_{AH} and τ_{WOSA} (p value = .2938). Even the test of median of Wilcoxon goes in the direction to not reject the hypothesis that the medians have significant differences (p-value = .5153). This may come as a surprise taking in consideration that both the procedures lose information, but WOSA loses data before 91 days from all classes while AH loses all data from one class but with no regard of time.

The *noise* in time series may be better understood under the light of a phenomenon that is observed adopting OSA from November 6th, 2011, a day when 10 subjects were *active* (see, Subject 4): among the remaining 50 inactive subjects, 28 came into activity ('newcomers') directly as rank 1st. Adopting WOSA instead of OSA, this ratio decreases into 23/50. In both cases, this phenomenon might be explained by a social mechanism of *optimization* reinforced by *herding* effect.

6 Conclusive Remarks

The substantial result from Sect. 5 is that removing information the impact of biases on the fairness of a crowdsourced ranking procedure is reduced. Weighting schemes, removing information from older 'reviews' have a better rationale than Ad hoc rating measures. Different *deltas* (i.d. the number of days before censoring) and weighting schemes might be worth to be tested in order to capture practical rules for crowd rating design.

Biased rating values in those newcomers which get into 1st rank (a 'capture-event') seem a phenomenon worth of further attention. An investigation over the value of the association over time between (A) capture-events in t and (B) The sign of $\tau(t-1) - \tau(t)$, may confirm the causal mechanism. However, further research on τ-statistic with different k, datasets and contexts is advisable to wide insights over crowd rating.

The main condition to satisfy the purpose of employment of τ-statistic is that the ranking should reflect a real competition among ranked subjects. This condition may be *fuzzy* or not well determined in real life, e.g. are movies competitors? Are a pizzeria and a sushi bar in the same area competitors? We think these questions get a positive answer only to a degree. However, a sushi bar in Tokio and one in Milan cannot be competitors (or at least, not from a economic perspective). Athletes running a marathon instead are doing a competition *by definition*. Although τ is always computable, we think the concept

of *rank-coherence* acquires proper sense when the social situation is well defined and accepted as a competition among participants: the lower the τ, the fairer is the ranking for the *public* and, in this sense, the competition is accountable.

A closing argument we desire to highlight is on the parametric measures to represent distribution of crowd rating. We argued that ordinal scales are the most common support, chosen with the aim to constrain scores. This reflects a historical heritage of 'what worked' in social research, at least according to our personal perception. An interesting alternative made possible through digitalisation of rating tools is a *switch* scale: a tool where a customer can move a kind of interactive digital lever, switching the score between all integers from 1 through 100. Although the format is technically similar of a Likert scale, the large amount of points in the scale could set a turning point in the debate about employment of mean or median for ordinal scales.

A combination of *switch* scales, generalized mixture models, weighting schemes and time series methods might be unified into a proper methodology of crowdsourced estimation of social phaenomena as suggested by authors mentioned (see, Sect. 2.1) in the debate on future developments of customer satisfaction methods.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

References

Aggarwal, C. C. (2016). *Recommender Systems*. Heidelberg: Springer.

Akerlof, G. A. (1970). The market for 'lemons': quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84*(3), 488–500.

Alvo, M., & Yu, P. L. H. (2014). *Statistical Methods for Ranking Data*. Heidelberg: Springer.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*(2), 130–147.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet Rating of Product Reviews. In M. Boughanem, et al. (Eds.), *ECIR 2009, LNCS 5478* (pp. 461–472). Berlin Heidelberg: Springer-Verlag.

Bai, T., Zhao, X., He, Y., Nie, J. Y., & Wen, J. R. (2018). Characterizing and predicting early reviewers for effective product marketing on e-commerce websites. *IEEE Transactions on Knowledge and Data Engineering, 30*(12), 1–14.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter, 9*(2), 75.

Clippinger, J. H. (2011). An inquiry into effective reputation and rating systems. In H. Masum & M. Tovey (Eds.), *The Reputation Society* (pp. 25–36). Cambridge MA: MIT Press.

Corain, L., Arboretti, R., & Bonnini, S. (2016). *Ranking of multivariate populations: A permutation approach with applications*. Boca Raton: CRC Press.

De Battisti, F., Nicolini, G., & Salini, S. (2010). The Rasch model in customer satisfaction survey data. *Quality Technology & Quantitative Management, 7*(1), 15–34.

Dellarocas, C. (2011). Designing reputation systems for the social web. In H. Masum & M. Tovey (Eds.), *The Reputation Society* (pp. 3–12). Cambridge: MIT Press.

Érdi, P. (2019). *Ranking*. The Unwritten Rules of the Social Game We All Play: Oxford University Press.
Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science, 38*(2), 189–200.
Farmer, R. (2011). Web reputation systems and the real world. In H. Masum & M. Tovey (Eds.), *The Reputation Society* (pp. 13–24). Cambridge (MA): MIT Press.
Fernández-Barcala, M., González-Díaz, M., & Prieto-Rodríguez, J. (2010). Hotel quality appraisal on the internet: a market for lemons? *Tourism Economics, 16*(2), 345–360.
Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: nature, purpose, and findings. *Journal of Marketing, 60*(4), 7–18.
Galton, F. (1907). *Vox Populi. Nature, 75,* 450–451.
Geiger, D., Schader, M., Rosemann, M., & Fielt, E. (2012). Crowdsourcing information systems - definition, typology, and design. *Proceeding of International Conference on Information Systems* (pp. 1–11), Orlando, FL.
Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210.
Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871.
Guttman, L. (1977). What is not what in statistics. *The Statistician, 26,* 81–107.
Iannario, M., & Piccolo, D. (2010). A new statistical model for the analysis of customer satisfaction. *Quality Technology & Quantitative Management, 7*(2), 149–168.
Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.
Jeacle, I., & Carter, C. (2011). In TripAdvisor we trust: rankings, calculative regimes and abstract systems. *Accounting, Organizations and Society, 36*(4/5), 293–309.
Kenett, R. S., & Salini, S. (2011). Modern analysis of customer satisfaction surveys: comparison of models and integrated analysis. *Applied Stochastic Models in Business and Industry, 27*(5), 465–475.
Khusro, S., Ali, Z., & Ullah, I. (2016). Recommender systems: Issues, challenges, and research opportunities. In K. Kim & N. Joukov (Eds.), *Information Science and Applications (ICISA) 2016* (pp. 1179–1189). New York: Springer.
Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM, 53*(4), 89–97.
Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 145–186). Boston: Springer.
Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol 1: Additive and Polynomial Representations*. San Diego, CA: Academic Press.
Langville, A., & Meyer, C. (2012). *Who's #1?: The Science of Rating and Ranking*. Princeton: Princeton University Press.
Soll J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning., Memory and Cognition*, *35*(3), 780–805.
Leal, F., Malheiro, B., & Burguillo, J. C. (2018). Analysis and prediction of hotel ratings from crowdsourced data. *WIREs Data Mining Knowledge Discovery, 9*(2), 1–9. https://doi.org/10.1002/widm.1296.
Lee, Y. J., Hosanagar, K., & Tan, Y. (2015). Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Science, 61*(9), 2241–2258.
Lewis, J. R. (1993). Multipoint Scales: mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction, 5*(4), 383–392.
Lewis, J. R., & Sauro, J. (2016). *Quantifying the User Experience: Practical Statistics for User Research*. Cambridge: Morgan Kaufmann.
Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1566–1576), Baltimore, MD.
Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.
Lucas, J. P., Luz, N., Moreno, M. N., Anacleto, R., Almeida Figueiredo, A., & Martins, C. (2013). A hybrid recommendation approach for a tourism system. *Expert Systems with Applications, 40*(9), 3532–3550.
Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95.
Mangel, M., & Samaniego, F. (1984). Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association, 79*(386), 259–267.
Mari L. & Ruffini, R. (2018). An analysis of Goodhart's law toward a shared conceptual framework of measurement across the sciences. Journal of Physics: Conference Series, 1065. doi: 10.1088/1742-6596/1065/7/072022
Melville, P., & Sindhwani, V. (2017). Recommender systems. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1056–1066). Berlin Heidelberg: Springer.

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Boston: Addison-Wesley.

Müller-Trede, J., Choshen-Hillel, S., Barneron, M., & Yaniv, I. (2018). The wisdom of crowds in matters of taste. *Management Science, 64*(4), 1779–1803.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In: *Proceedings of the 21st international conference on World Wide Web* (pp. 201–210). Lyon.

Piccolo, D., & D'Elia, A. (2008). A new approach for modelling consumers' preferences. *Food Quality and Preference, 19*(3), 247–259.

Piccolo, D., & Simone, R. (2019). Rejoinder to the discussion of "The class of cub models: statistical foundations, inferential issues and empirical evidence". *Statistical Methods and Applications, 28*(3), 389–435.

Pizam, A., Shapoval, V., & Ellis, T. (2016). Customer satisfaction and its measurement in hospitality enterprises: a revisit and update. *International Journal of Contemporary Hospitality Management, 28*(1), 2–35.

Proietti, T. (2019). Discussion of "The class of CUB models: statistical foundations, inferential issues and empirical evidence". *Statistical Methods and Applications, 28*(3), 451–456.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321–333, Berkeley, CA.

Salganik, M. J. (2018). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5792), 854–856.

Si, M., & Li, Q. (2020). Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review, 53,* 291–319.

Sinai, Y. G. (1976). *Introduction to Ergodic Theory*. Princeton: Princeton University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: heuristics and Biases. *Science, 185,* 1124–1131.

Varian, H. R. (2016). The economics of Internet search. In J. Bauer & M. Latzer (Eds.), *Handbook on the Economics of the Internet* (pp. 385–394). Cheltenham: Edward Elgar Publishing.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47*(1), 65–72.

Wallis, K. F. (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science, 29*(3), 420–424.

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing. *Proceedings of the VLDB Endowment, 10*(5), 541–552.