SORT
Statistics and Operations Research Transactions

Volume 42
Number 2
July-December 2018
ISSN: 1696-2281
eISSN: 2013-8830

www.idescat.cat/sort/

9 771696 228009 42002

SORT Volume 42 (2) July-December 2018

SORT

Statistics and Operations Research Transactions

Volume 42
Number 2, July-December 2018
ISSN: 1696-2281
eISSN: 2013-8830

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Calidad de Revistas Científicas Españolas
FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2016

ISSN: 1696-2281
eISSN: 2013–8830
SORT 42(2) July-December (2018)

SORT

Statistics and Operations Research Transactions

Coediting institutions
Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions
Spanish Region of the International Biometric Society
Societat Catalana d'Estadística

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 42
Number 2
July-December 2018
ISSN: 1696-2281
eISSN: 2013-8830

In memoriam: Heinz Neudecker (1933-2017)

Invited article

Articles

Selected articles from *XVI Conferencia Española de Biometría 2017*

In memoriam: Heinz Neudecker (1933-2017)

Heinz Neudecker was born on October 3, 1933 in The Hague. His parents were Austrian. They moved to The Netherlands in the early 1930s. He married Erica Engels in 1961.

Neudecker obtained his Master of Science in Rotterdam, 1959. He earned his Ph.D. after completing his research in Birmingham, 1967. In agreement with his leftist ideology, the subject of the thesis was investment criteria in Soviet economy planning. However, his vocation in linear algebra finally dominated and his thesis was "Matrix Methods for Econometric Research", a topic which remained with him throughout his career. He made fundamental contributions in this area, which had a deep impact in the theory and practice of multivariate analysis, econometrics, and related disciplines. After having worked in several economic research institutions and universities, in Birmingham, Ankara, and Brussels, he was appointed Professor at the Faculty of Economic Science and Econometrics of the University of Amsterdam, where he remained for the rest of his career. After retirement, he remained at that institution as Emeritus Professor of Mathematics, Mathematical Economics, Econometrics and Statistics.

My first contact with Neudecker was during the Second Catalan International Symposium on Statistics (Barcelona, September 1986). Later, I met him at the University of Amsterdam (July 1988). Again I met Neudecker in Barcelona in 1989. Then he published two papers in **Qüestiió**. Later he gave some seminars at the University of Barcelona (May 1993). His main activity took place at the Universitat Pompeu Fabra. At the end of the 1980s, through Albert Satorra, he consolidated the contact with the Catalan community of statisticians. He was a frequent visitor of the Universitat Pompeu Fabra where he collaborated with Satorra. Presumably, the first research paper published in an international journal with signature "Universitat Pompeu Fabra" was the paper Neudecker and Satorra "LISREL: Gradient and Hessian of the fitting function", *Statistics and Probability Letters,* 11, 57-61, 1991.

Neudecker was a man of action, an architect, for research contacts across European campuses, reaching remote points such as the University of Tartu (Estonia). This was not only in research, but also for students' mobility. Neudecker probably started the first program in cooperation with the Baltic republics, joining the Universitat Pompeu Fabra with the University of Amsterdam, and the University of Tartu. The connection with Tartu propitiated new contacts with researchers on multivariate analysis, for example Tonu Kollo and Ene-Margit Tiit, and the attendance to several international conferences on multivariate statistics in Tartu (Estonia), where we came into contact with prominent statisticians of around the world. In this way Neudecker impacted on the internationality of the multivariate analysis research groups in Barcelona.

On my personal anecdotic memory, Neudecker spent some days in my summer house (near Tossa de Mar) in August 1990. One afternoon he walked along the rather labyrinthian resort, going alone and taking a map. He got lost and two guardians, commanded by a strict German manager of the resort, "arrested" him. After suitable explanations, Heinz was identified and returned home. Then he recalled the flight from Austria of his parents, who walked a long distance sixty years before, probably because of the forthcoming Anschluss.

Neudecker was a man of quick decisions. He had a passion for the Mediterranean life and culture but hated the beach tourism. He acquired a house in the old town of Sant Fruitós de Bages

(near Barcelona). There he had a terrace with certainly the best panoramic view of Montserrat, with a changing panorama of colors along the day. With his wife Erica, they converted this terrace as the best place for producing matrix algebra and multivariate analysis research. In between research pauses, Heinz and Erica explored all the rivers, natural fountains and paths in the environs of Sant Fruitós, becoming more experts on the natural resources of the town than other neighbours. Some of the areas they explored and enjoyed came to be destroyed later.

I visited his house in Sant Fruitós several times. Since I was doing research on multivariate analysis, Heinz connected me (and others, as it is said above), with researchers of such as Tonu Kollo and Ene-Margit Tiit, and I attended several conferences on multivariate statistics in Tartu (Estonia). In 1990 I attended a conference in Rome on distributions with given marginals, which I also told Tiit about, and she also attended this conference. Both Kollo and Tiit gave seminars at the University of Barcelona in September, 1991.

With more than 139 publications indexed in the ISI web of knowledge, Neudecker reshaped the field of matrix algebra for econometrics and multivariate analysis. He is the author (with one of his Ph.D. students, Jan Magnus) of the most influential book in the area, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, with first edition in 1988 and a revised edition in 1991. The second edition was published in 1999.

Most Neudecker's papers (alone or with Magnus, Satorra, Kollo, Styan, Liu, Van de Velden and others), were interesting and seminal. He studied the commutation matrix, the Kronecker and Hadamard matrix products, high-order matrix derivatives, inequalities, covariances and asymptotic properties of eigenvalues. In five papers published in **Qüestiió**, he also studied the matrix Haffian, derivations of specific matrices and a new proof of the Milliken-Akdeniz theorem. He proposed several problems in the Teaching Section and sent solutions to problems of other authors. Heinz also participated in the discussion about the new title of this journal, **SORT**. During the period 1997-2017 Neudecker was Associate Editor of **Qüestiió** and belonged to the Editorial Advisory Committee of **SORT**. His results on Haffian matrices are a good patrimony of our Catalan journal. Indeed, the investigations carried out by Neudecker, most published in prestigious journals, constitute a real advance for the science.

To acknowledge Neudecker's contributions to matrix calculus, statistics and econometrics, in 2000 Heijmans, Pollock and Satorra edited *Innovations in Multivariate Statistical Analysis*. Many authors published chapters related to the topics investigated by Neudecker. Since one topic was correspondence analysis, I contributed (with Fortiana and Greenacre) with a continuous generalization of correspondence analysis.

Heinz Neudecker passed away on December 5, 2017.

Carles M. Cuadras, Past Editor in Chief of SORT.
Barcelona, December 2018.

SORT 42 (2) July-December 2018, 101-124 DOI: 10.2436/20.8080.02.71

Evidence functions: a compositional approach to information

J.J. Egozcue[1] and V. Pawlowsky-Glahn[2]

Abstract

The discrete case of Bayes' formula is considered the paradigm of information acquisition. Prior and posterior probability functions, as well as likelihood functions, called evidence functions, are compositions following the Aitchison geometry of the simplex, and have thus vector character. Bayes' formula becomes a vector addition. The Aitchison norm of an evidence function is introduced as a scalar measurement of information. A fictitious fire scenario serves as illustration. Two different inspections of affected houses are considered. Two questions are addressed: (a) which is the information provided by the outcomes of inspections, and (b) which is the most informative inspection.

MSC: 60A10, 60E10, 62E10

Keywords: Evidence function, Bayes' formula, Aitchison geometry, compositions, orthonormal basis, simplex, scalar information

1. Introduction

Each summer fires in forests and suburban areas affect houses, industries, and the whole environment. When this occurs, authorities need to get a quick diagnostic of damages, both for mitigation of effects, evaluation of economic costs and, especially, for evacuation of population from houses and planning of further actions. Airborne photography and visual inspection of houses are emergency means to classify houses into categories, usually corresponding to (a) buildings that can be reoccupied by the previously evacuated people, (b) buildings that require some repairs, (c) buildings that are largely damaged or (d) buildings that are collapsed. The impact of such diagnostics is critical, as the damnified population can or cannot recover their homes, do or do not receive economic compensations, depending on the result of the inspection. Typical questions are: How uncertain/informative are the results of an inspection? Which type of inspection is more reliable? What is the amount of information after inspections? These questions are

[1] Dept. Civil and Environmental Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain.
E-mail: juan.jose.egozcue@upc.edu

[2] Dept. of Computer Sciences, Applied Mathematics, and Statistics, Universitat de Girona, Spain.
E-mail: vera.pawlowsky@udg.edu

Received: August 2018

related to the quantification of information provided by an experiment (inspections) and, therefore, should be answered by the statistical theory of information.

The above scenario of fires is not the only one where the questions on information provided by experiments are relevant. A very similar situation corresponds to many hazardous situations like earthquakes, floods, hurricanes, terrorist attacks... Also, clinic diagnostic of deseases, military actions or, in general, operational decisions under uncertainty correspond to the same type of scenario, which can be modelled as a collection of uncertain states or events, frequently assumed non-overlapping, to which some prior probabilities describing uncertainty on the true event are assigned; then, one or more experiments (diagnostic tests, inspections) are carried out, trying to reduce uncertainty; finally, after the results of the experiments, the updating of the probabilities (posterior probabilities) may allow to use the information available in decision making schemes. This scheme has been well known for decades, and still maintains its validity (e.g. Benjamin and Cornell, 1960).

The previous questions have been addressed from different points of view in information theory, specially following the line proposed by Lindley (1956). However, information theory was born from the study of coding and communication (Shannon, 1948, Shannon and Weaver, 1949, McMillan, 1953) and built on an early contribution by Hartley (1928), where logarithms of probabilities were identified as a measure of information. The initial development of the theory in the framework of communications and its particular syntaxis may be the reason why the statistical theory of information was developed some years later (e.g. Kullback and Leibler, 1951a, Kullback, 1997, Lindley, 1956, Khinchin, 1957, Ash, 1990). In medicine, diagnostic tests were studied, for instance, by Aitchison and Kay (1975) (see also Aitchison, Kay and Lauder, 2005).

The statistical theory of information is directly related to the concept of entropy. This is viewed as an average of measures of uncertainty (Shannon, 1948, McMillan, 1953) which is common to all branches of information theory. More rarely, information acquisition is linked to the Bayes' formula (Lindley, 1956) and its extensions, for example Dempster's rule in the theory of beliefs (Yager and Liu, 2008).

The aim of the present contribution is rethinking the bases of information theory from the point of view of compositional data analysis (Aitchison, 1986, Pawlowsky-Glahn and Buccianti, 2011, Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). For completeness, Appendix A is a summary of Aitchison geometry for compositions, introducing notation and basic tools used. The main proposal is that information is a vector magnitude identified as a composition. These compositions are here called evidence functions, e-functions for short, and include traditional (discrete) probability functions and also likelihood functions. The Aitchison norm of e-functions as compositions (see Appendix A) is used as a scalar measure of information called e-information. This is in contrast to Shannon information and its related magnitudes, which were developed as scalar measures of information. Other points which are relevant to this proposal are:

- The Bayes' formula (discrete case) is the paradigm of information acquisition;

- The Bayes' formula is a vector additive Abelian group operation in the simplex endowed with the Aitchison geometry;
- Discrete probability functions (prior, posterior) and discrete likelihood functions are compositions and, consequently, they share the same properties.

Section 2 reviews concepts of compositional geometry and identifies evidence functions involved in Bayesian updating as compositions (see also Appendix A). Section 3 introduces a scalar measure of information, namely the Aitchison norm of an evidence function. Its properties characterize it as a proper measure of information. Section 4 discusses the acquisition of information through a fictitious fire scenario and inspections of affected houses.

2. Bayes theorem, evidence functions and compositions

Consider the fire scenario in which a number of isolated, but close, houses have been affected. It is assumed that these houses can be in $D = 4$ states, denoted A_i, $i = 1,2,\ldots,D$, which can be identified with *service* or *no damage* (Nod), *moderate damage* (Mod), *severe damage* (Sev) and *ruin or collapse* (Col). These states are assumed non-overlapping. Based on previous urban studies, there is a perception that, after the fire, most houses will remain in service (80%) or with little damage (15%), meanwhile some of them will be largely damaged (4%) or in ruin (1%). In the Bayesian terminology, the vector of probabilities $\mathbf{p} = (p_1,\ldots,p_D) = (0.80, 0.15, 0.04, 0.01)$, is known as prior or initial probabilities (this prior is reported in Table 1 as $\mathbf{p}^{(1)}$). The vector $\mathbf{p}$ is a composition. In fact, expressed as proportions or as percentages, the information is exactly the same; in particular, ratios between components remain the same. Moreover, the set of odds obtained by the ratios between components contains all the relative information and could be used to retrieve the numerical value of $\mathbf{p}$. These simple features characterize $\mathbf{p}$ as a D-part composition. The fact that the relative information contained in $\mathbf{p}$ remains unaltered when it is multiplied by a positive constant corresponds to the *scale invariance principle* of compositional data, and to its consequence, namely that the relative information is provided by the ratios of components (Aitchison, 1986, 1994). More recently, compositional equivalence has been defined as the condition that vectors of positive components which are proportional are compositionally equivalent (Barceló-Vidal and Martín-Fernández, 2016, Pawlowsky-Glahn et al., 2015, Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 2001). The generated equivalence classes can always be represented in a unitary D-part simplex, denoted $\mathbb{S}^D$, so that the sum of the parts is one, as in the usual normalization of probability. For simplicity, the projection of a non-normalized composition onto $\mathbb{S}^D$ is denoted by the closure operator $\mathscr{C}$.

Frequently, only some parts of the vector $\mathbf{p}$ are considered. For instance, only reparable buildings, i.e. only the three first parts, are taken into account. This restriction is

a *subcomposition*. A subcomposition like $\mathscr{C}(p_1, p_2, p_3)$ corresponds to a conditional probability vector $(p_1/p_c, p_2/p_c, p_3/p_c)$ with $p_c = p_1 + p_2 + p_3$. This suggests that the identification of vectors of probabilities with compositions is natural.

Returning to the fire scenario, assume that a visual inspection of affected houses has been devised. The inspectors, after a quick visit of a building, decide to assign a color code according to their perception: green for service or no damage, orange for moderate damage, red for severe damage, and black for ruin or collapse. Obviously, this kind of assessment is quite uncertain, and the color codes do not correspond exactly to the real state of the building. Let R be the result of an inspection (e.g. orange: moderate damage in a visual inspection). For each possible result R, the conditional probabilities $q_i = \Pr(R|A_i)$, $i = 1, 2, \ldots, D$, characterize the experiment. In fact, the likelihood function associated with R, $\mathbf{q} = (q_1, q_2, \ldots, q_D)$, allows to apply Bayes' formula to obtain final or posterior probabilities $\mathbf{f} = (f_1, f_2, \ldots, f_D)$ as

$$\mathbf{f} = C \cdot (p_1 q_1, p_2 q_2, \ldots, p_D q_D), \quad C = \frac{1}{\Pr(R)} = \left(\sum_{k=1}^{D} p_k q_k \right)^{-1}, \tag{1}$$

with $\mathbf{p}$ the vector of prior probabilities. This expression of the final probabilities, after the observation of R, matches exactly the definition of perturbation in the simplex, as pointed out by Aitchison (1986). Perturbation is an Abelian group operation in the simplex, and it is the addition in the Aitchison geometry for compositions (Pawlowsky-Glahn and Egozcue, 2001, Pawlowsky-Glahn et al., 2015), that is, Bayesian updating is a shift of the prior probabilities to the final probabilities by the likelihood. The simplex $\mathbb{S}^D$, endowed with perturbation ($\oplus$, group operation), powering ($\odot$, external multiplication) and Aitchison inner product, is a $(D-1)$-dimensional Euclidean space (Billheimer, Guttorp and Fagan, 2001, Pawlowsky-Glahn and Egozcue, 2001) (see Appendix A for detailed definitions). Therefore, denoting perturbation by $\oplus$, the Bayes formula is simply

$$\mathbf{f} = \mathbf{p} \oplus \mathbf{q}, \tag{2}$$

where no reference to the normalizing constant is necessary due to the compositional equivalence. Commonly, it is assumed that the difference between vectors of probabilities (initial or prior, final or posterior) and the likelihood function is that the latter is not normalized. The three symbols $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{f}$ are considered as compositions: in fact, the normalization of probabilities is irrelevant and the *likelihood principle* (Birnbaum, 1962), preconizes equal inferences for proportional likelihood functions, thus the likelihood itself is a composition.

The standard information theory (e.g. Gray, 2011), assigns a measure of uncertainty to a vector of probabilities called (Shannon) *entropy*,

$$\mathscr{H}_S(\mathbf{p}) = -\sum_{i=1}^{D} p_i \log p_i \,. \tag{3}$$

The terms $\log(1/p_i)$, $i = 1,2\ldots,D$, were proposed by Hartley (1928) as information provided by the observation of the event A_i. Defining a random variable which takes the values $\log(1/p_i)$ with probability p_i, Equation 3 is the mean of such random variable. Then, within the framework of the standard information theory, differences of entropies, for instance, after and before observing the result of an experiment, gives a measure of information. There are several ways of measuring these differences of uncertainties or entropies. The most popular is the Kullback-Leibler divergence (Kullback, 1997) which considers the differences $\log(1/f_i) - \log(1/p_i)$ and takes the mean using the posterior probabilities f_i

$$\mathscr{I}_{KL}(\mathbf{f}:\mathbf{p}) = \sum_{i=1}^{D} f_i \log \frac{f_i}{p_i} \ ,$$

using the notation $\mathbf{p}$ (prior) and $\mathbf{f}$ (final) in the Bayes' formula (1). Following Lindley (1956), the information, assigned to a vector of probabilities like $\mathbf{p}$, is

$$\mathscr{I}_S(\mathbf{p}) = \sum_{i=1}^{D} p_i \ \log p_i = -\mathscr{H}_S(\mathbf{p}) \ . \tag{4}$$

These measures of information, and many other entropy divergences (e.g. Martín-Fernández, 2001, and references therein) are not invariant under scaling of $\mathbf{p}$ and $\mathbf{f}$ and, therefore, the computation of $\mathscr{I}_S$ or $\mathscr{I}_{KL}$ requires that $\mathbf{p}$ and $\mathbf{f}$ are normalized, i.e. their components sum to 1. This is a major inconvenience for likelihood functions which, in general, are not normalized. A symmetrized and compositional version of the Kullback-Leibler divergence is given by Martín-Fernández (2001).

From the compositional point of view, the three compositions, $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{f}$, live in the same space, $\mathbb{S}^D$, equipped with the Aitchison geometry (see discussion in the continuous case by Egozcue et al., 2013). Furthermore, the three compositions model the uncertainty on the actual event A_i or, from the opposite point of view, the evidence in favour of these events. This motivates calling the three compositions *evidence functions* or e-functions for short.

With this terminology, evidence functions are vectors and Bayes updating is just vector addition (perturbation) in the space of e-functions. Figure 1, illustrates these facts. In the left panel the three evidence functions (prior, likelihood and posterior) are represented as probabilities. The likelihood corresponds to the visual observation of moderate damage (vMod), the prior corresponds to the subjective impression of almost complete destruction of houses in the neighbourhood ($\Pr(A_4) = 0.7$, $\Pr(A_1) = \Pr(A_2) = \Pr(A_3) = 0.1$), which was selected for clarity of the picture. The right panel shows the three evidence functions as vectors, in which the posterior is the vector sum of the prior and the likelihood. The simplicity of the vectorial representation contrasts with the difficulties in comparing the proportions in the left panel.

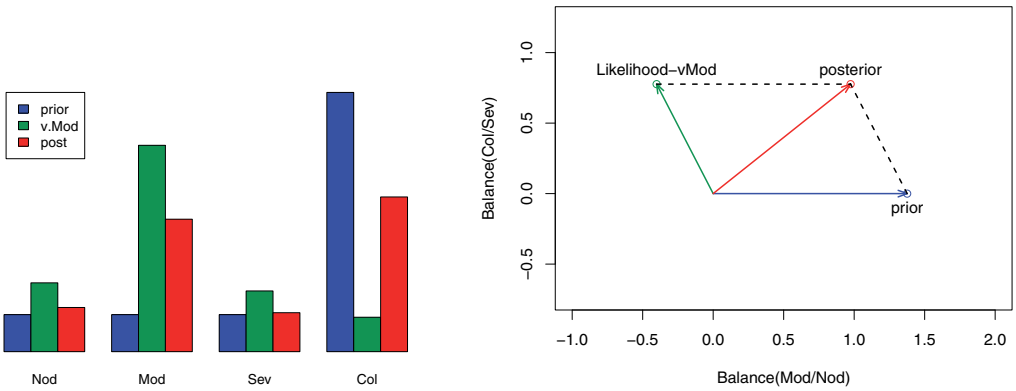


Figure 1: *Left panel: evidence functions, prior (blue), likelihood, corresponding to R =vMod (green), posterior (red) for the actual states Nod, Mod, Sev, Col. Right panel: the Bayes' formula in the two first coordinates; it appears as a vector addition. See definition of coordinates in Section 4.*

The consequences of the vectorial character of evidence functions are multiple. Bayes' formula, (1) and (2), has the equivalent expression in ilr coordinates or in clr coefficients (see Appendix A), that is

$$\mathrm{ilr}(\mathbf{f}) = \mathrm{ilr}(\mathbf{q}) + \mathrm{ilr}(\mathbf{p}) \quad , \quad \mathrm{clr}(\mathbf{f}) = \mathrm{clr}(\mathbf{q}) + \mathrm{clr}(\mathbf{p}) \ ,$$

where the additive character of the Bayes updating is explicit. The size of a vector is described by its norm (or a monotone function of it), regardless of its direction, a fact which motivates the definition of a scalar measure of information (Section 3). Vectors in a Euclidean space can be parallel, orthogonal, unitary; they can be projected one onto other, approximated by linear combinations of other vectors; distances between them are available, they can be expressed in coordinates. Remarkably, all these concepts and operations can be applied to or performed on evidence functions and, consequently, to information: *information* represented by *evidence functions* is a vectorial magnitude.

The parallelogram property of vectors in Euclidean spaces can be rephrased in terms of Bayesian updating. Consider the result of an experiment which provides a likelihood function $\mathbf{q}$. Imagine that two different priors, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, are proposed, for instance, in the fire scenario the prior initially mentioned, $\mathbf{p}^{(1)} = (p_1^{(1)}, \ldots, p_D^{(1)}) = (0.80, 0.15, 0.04, 0.01)$, and that used in Figure 1, denoted $\mathbf{p}^{(2)} = (p_1^{(2)}, \ldots, p_D^{(2)}) = (0.1, 0.1, 0.1, 0.7)$ (Table 1). The Aitchison distance between $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, $\mathrm{d}_a(\mathbf{p}^{(1)}, \mathbf{p}^{(2)})$, can be easily computed using any of the expressions in Equation (12) of Appendix A. In the example, this Aitchison distance is approximately 4.65 and the norms are $\|\mathbf{p}^{(1)}\|_a = 3.24$ and $\|\mathbf{p}^{(2)}\|_a = 1.69$, that is, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are neither orthogonal nor parallel (see Table 1). In fact, the two priors were designed to represent very different situations: $\mathbf{p}^{(1)}$ assumes that the zone, being largely affected by fire, has not been completely destroyed; for $\mathbf{p}^{(2)}$ houses which are completely destroyed are a large majority. These two priors $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$ can be updated with the same likelihood $\mathbf{q}$, thus obtaining two different final probabil-

ities $\mathbf{f}^{(1)} = \mathbf{p}^{(1)} \oplus \mathbf{q}$, $\mathbf{f}^{(2)} = \mathbf{p}^{(2)} \oplus \mathbf{q}$. Elementary properties of Aitchison geometry, as a Euclidean geometry, state that the perturbation difference between $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ is that of the priors, that is

$$\mathbf{p}^{(1)} \ominus \mathbf{p}^{(2)} = (\mathbf{p}^{(1)} \oplus \mathbf{q}) \ominus (\mathbf{p}^{(2)} \oplus \mathbf{q}) = \mathbf{f}^{(1)} \ominus \mathbf{f}^{(2)} \,.$$

Hence, due to the parallelogram property of vectors in Euclidean spaces, the relation between the Aitchison distances is

$$\mathrm{d}_a(\mathbf{p}^{(1)}, \mathbf{p}^{(2)}) = \mathrm{d}_a(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}) = \mathrm{d}_a(\mathbf{p}^{(1)} \oplus \mathbf{q}, \mathbf{p}^{(2)} \oplus \mathbf{q}) \,.$$

This means that the difference of prior e-functions is equal to the difference of posterior e-functions, provided that the likelihood was the same in the application of Bayes' formula. This result is different when using Shannon information or Kullback-Leibler divergence, for which the information provided by an experiment depends on the prior, a property which is well accepted in Bayesian statistics. These facts, are illustrated in Section 4.

3. Scalar information in an evidence function

Which of two results of an experiment is the most informative? This natural question cannot be answered if information is a vector, as real vectors cannot be ordered. A scalar measure of information associated with e-functions is needed, despite their vectorial character. The norm of an e-function, as a composition represented in $\mathbb{S}^D$, is the natural candidate for a scalar measure of information. Consequently, the scalar information contained in an e-function, $\mathbf{f} = (f_1, f_2, \ldots, f_D) \in \mathbb{S}^D$, is defined as

$$\mathscr{I}_e(\mathbf{f}) = \|\mathbf{f}\|_a \,, \tag{5}$$

where $\|\cdot\|_a$ denotes the Aitchison norm of a composition (Appendix A, Eq. 13). Therefore, the scalar information $\mathscr{I}_e$ has all standard properties of a vector norm. Some properties, which have a meaningful interpretation in the framework of information, are detailed below (Egozcue and Pawlowsky-Glahn, 2011). It is worth comparing the following properties of $\mathscr{I}_e$ with those which are satisfied by the Shannon entropy, $\mathscr{H}_S$, for instance, those proposed by Shannon (1948), Khinchin (1957) or Ash (1990). Entropy is conceived as a measure of uncertainty, and information is then defined from differences between initial and final entropy (Kullback and Leibler, 1951a, Khinchin, 1957, Ash, 1990), or even as negative entropy (Eq. 4) (Lindley, 1956).

Null e-Information. A flat e-function $\mathbf{n} = (1/D, 1/D, \ldots, 1/D)$ does not provide any information, as $\mathscr{I}_e(\mathbf{n}) = 0$ corresponds to the neutral element in $\mathbb{S}^D$. This property is shared by all definitions of measures of information, alternatively entropy. Note that,

reciprocally, for any e-function, $\mathscr{I}_e(\mathbf{f}) = 0$ implies $\mathbf{f} = \mathbf{n}$. This is due to the fact that $\|\mathbf{f}\|_a$ is the Aitchison distance (not a divergence) from $\mathbf{f}$ to the neutral element $\mathbf{n}$.

Continuity. Also common to all definitions of information is the continuity of information/entropy with respect to each component of the e-function. The ilr coordinates (Eq. 10 in Appendix A) are continuous functions of the components of the e-functions. Also the Aitchison norm (Eq. 13 in Appendix A) is a continuous function of the ilr coordinates. Then, $\mathscr{I}_e(\mathbf{f})$ is a continuous function of the $\mathbf{f}$ components. The only critical points are those in which one or more parts $f_i = 0$, as null components place the value of $\mathscr{I}_e(\mathbf{f})$ at infinity. Knowledge of the impossibility of event A_i represents the strongest information. It forces the change of sample space just by removing event A_i. This is opposite to the case of Shannon information, where $-\log f_i$ is minus infinity before averaging, while $\mathscr{I}_S$ (Eq. 4) remains unaltered after averaging with null probability.

Monotonicity. A set of properties was used to introduce (Shannon) *entropy*, $\mathscr{H}_S$, in an axiomatic way. Simultaneously, entropy was taken as opposite to information (for the Shannon case $\mathscr{I}_S = -\mathscr{H}_S$). Following Ash (1990), the monotonicity property for Shannon entropy $\mathscr{H}_S$ is that, if d and D, $d < D$, are the number of parts of two neutral compositions then

$$\mathscr{H}_S(1/d, 1/d, \ldots, 1/d) < \mathscr{H}_S(1/D, 1/D, \ldots, 1/D) \ ,$$

which, loosely speaking, means that uncertainty or entropy increases with the number of components, here written for neutral compositions. This statement is not really useful for a measure of information which attains a null value at neutral elements, like $\mathscr{I}_e(1/d, 1/d, \ldots, 1/d) = 0$. In the case of $\mathscr{I}_e$ this kind of monotonicity is captured by the subcompositional dominance property of the Aitchison distance (e.g. Aitchison, 1983, Egozcue and Pawlowsky-Glahn, 2018), which is formulated as follows. Let $\mathbf{x}$ and $\mathbf{y}$ be compositions in $\mathbb{S}^D$ and their corresponding d-part ($d < D$) subcompositions $\mathbf{x}_d$ and $\mathbf{y}_d$. Then, $\mathrm{d}_a(\mathbf{x}, \mathbf{y}) \geq \mathrm{d}_a(\mathbf{x}_d, \mathbf{y}_d)$. When $\mathbf{y} = (1/D, \ldots, 1/D)$ (the neutral element), distances become norms and

$$\mathscr{I}_e(\mathbf{x}) \geq \mathscr{I}_e(\mathbf{x}_d) \quad , \quad D > d \ ,$$

which means that the information contained in a d-part subcomposition of an e-function is always less than or equal to the information contained in the (D-part) original e-function.

Null information extension. In Shannon entropy/information theory, extending the e-function with zeroes does not decrease entropy or increase information (e.g. Khinchin, 1957). This is a direct consequence of the fact that, for $p_{D+1} = 0$ the term $p_{D+1} \ln p_{D+1}$ is assumed null, and the previous information in Equation (4) remains unaltered after adding the term. This situation is completely different for $\mathscr{I}_e$. It can be proven that

$$\mathscr{I}_e(\mathbf{x}) = \mathscr{I}_e(\mathbf{x}, x_{D+1}) \quad \text{if and only if} \quad x_{D+1} = \mathrm{g_m}(\mathbf{x}) \ ,$$

that is, adding a part, x_{D+1}, equal to the geometric mean of the previous e-function does not alter $\mathscr{I}_e$. In fact, the extended composition can be represented using a system of ilr coordinates valid for $\mathbf{x}$, plus a new coordinate

$$b_D = \sqrt{\frac{D}{D+1}} \ \log \frac{x_{D+1}}{\mathrm{g_m}(\mathbf{x})} = 0 \ ,$$

which corresponds to completing a previous Sequential Binary Partition (SBP) (see Appendix A) for $\mathbf{x}$ with a sign code row $(-1,-1,\ldots,-1,+1)$. When computing the square Aitchison norm of $(\mathbf{x}, x_{D+1})$ a null term (Eq. 13 in Appendix A) is added.

The idea that extending a likelihood function, or other e-function, with zeros does not change the information provided by the experiment is counterintuitive: the result of the experiment informs the analyst that one or more categories are impossible, which would imply a great amount of information (infinite if using $\mathscr{I}_e$ as an information measure). This null extension seems acceptable when speaking of entropy or uncertainty: adding a null probability term to the e-function does not increase uncertainty. This reveals that Shannon entropy, $\mathscr{H}_S$ should have a more elaborated relation with information than just that expressed by $\mathscr{I}_S = -\mathscr{H}_S$; this can be seen in alternative interpretations of both magnitudes (Kullback and Leibler, 1951b, Ash, 1990).

Decomposition of an e-function. Consider a D-part e-function, $\mathbf{y}$, built appending two compositions, $\mathbf{x}_1$ with D_1 parts and $\mathbf{x}_2$ with D_2 parts. Then, $D = D_1 + D_2$. The compositions are appended after multiplying by arbitrary positive constants a_1 and a_2; that is, $\mathbf{y} = (a_1\mathbf{x}_1, a_2\mathbf{x}_2)$. The information conveyed by $\mathbf{y}$ is then

$$\mathscr{I}_e^2(\mathbf{y}) = \mathscr{I}_e^2(\mathbf{x}_1) + \mathscr{I}_e^2(\mathbf{x}_2) + \frac{D_1 D_2}{D_1 + D_2} \log^2 \frac{a_1 \mathrm{g_m}(\mathbf{x}_1)}{a_2 \mathrm{g_m}(\mathbf{x}_2)} \ . \tag{6}$$

The role of a_1 and a_2 is quite irrelevant, but they highlight the possibility of renormalizing the two compositions.

This kind of property differs from the corresponding property of Shannon entropy, mainly due to the assumed scalar character of information, and also to the need of renormalization. The property for the Shannon entropy, known as grouping axiom (Ash, 1990), is

$$\mathscr{H}_S(\mathscr{C}\mathbf{y}) = m(\mathbf{x}_1)\mathscr{H}_S(\mathscr{C}\mathbf{x}_1) + m(\mathbf{x}_2)\mathscr{H}_S(\mathscr{C}\mathbf{x}_2) + \mathscr{H}_S(m(\mathbf{x}_1), m(\mathbf{x}_2)) \ ,$$

where $\mathscr{C}$ is the closure operation (Appendix A), and $m(\mathbf{x}_k)$ is the sum of the components of $\mathscr{C}\mathbf{y}$ within the composition $\mathbf{x}_k$ ($k = 1,2$). Note that the computation of $\mathscr{H}_S$ requires normalization, and $m(\mathbf{x}_k)$ ($k = 1,2$) are the dividing normalization constants.

Independent probability table. Let be $\mathbf{x}_1 \in \mathbb{S}^{D_1}$ and $\mathbf{x}_2 \in \mathbb{S}^{D_2}$ two e-functions and $A = [a_{ij}]$ a (D_1, D_2) table of probabilities such that $a_{ij} = x_{1i}x_{2j}$. Then A, up to normalization,

is an independent table of probabilities. The scalar information associated with this table as e-function is

$$\mathscr{I}_e^2(A) = D_2\ \mathscr{I}_e^2(\mathbf{x}_1) + D_1\ \mathscr{I}_e^2(\mathbf{x}_2)\ . \tag{7}$$

To prove this statement, construct a (D_1, D_2) table A_2 with D_1 identical rows, each one equal to $\mathbf{x}_2$. Similarly, build a (D_1, D_2) table A_1 with D_2 identical columns, each one equal to $\mathbf{x}_1$. The entry-wise multiplication, or matrix perturbation $A_1 \oplus A_2$ (Egozcue et al., 2015), of these two tables is A. In Egozcue et al. (2015) it is proven that A_1 and A_2 as compositions are orthogonal, $\langle A_1, A_2 \rangle_a = 0$. Consequently, the square Aitchison norm of A is the sum of the square Aitchison norms of A_1 and A_2 (Pythagoras' theorem). On the other hand, the square norm $\|A_1\|_a^2 = D_2 \|\mathbf{x}_1\|_a^2$, as proven by Egozcue and Pawlowsky-Glahn (2019, Appendix A). A similar result holds for $\|A_2\|_a^2$, what implies the statement.

This is not what is expected in the Shannon information theory, in which the result is $\mathscr{H}_S(A) = \mathscr{H}_S(\mathbf{x}_1) + \mathscr{H}_S(\mathbf{x}_2)$, as reported, for instance, by Shannon (1948). The main difference with respect to Equation (7) is that additivity of entropy or information is thought in a scalar form in the Shannon theory; in the compositional approach information is thought as a vector (composition). In this case, independence is translated into orthogonality, thus reproducing the Pythagorean sum of squares in a Euclidean space.

Unit of information in evidence functions. The *bit* has been accepted as a unit of information since early works in the field. A bit is the Shannon information unit (using logarithms in basis 2) conveyed by an equiprobable binary code. It is obvious that this kind of definition is well adapted to the study of communications and coding theory. However, it is almost not interpretable in the present context of evidence functions and the scalar measure of information $\mathscr{I}_e$. In its place, a new unit of information adapted to e-functions is here proposed.

Consider an e-function $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ and a perturbation with a non closed composition $\mathbf{q} = (u, u^{-1}, 1, 1, \ldots, 1)$, $u = \exp(\sqrt{1/2})$. Then, $\mathbf{f} = \mathbf{p} \oplus \mathbf{q}$ is a shift of $\mathbf{p}$ towards $\mathbf{f}$. In order to compute $\mathscr{I}_e(\mathbf{q})$, one can decompose $\mathbf{q}$ into (u, u^{-1}) and the neutral element $\mathbf{n}$, and use the decomposition property (6) which yields

$$\mathscr{I}_e^2(\mathbf{q}) = \mathscr{I}_e^2(u, u^{-1}) = \left(\sqrt{\frac{1}{2}}\ \log \frac{u}{1/u} \right)^2 = 1\ .$$

Therefore, the perturbing composition $\mathbf{q}$ has a unit e-information, $\mathscr{I}_e(\mathbf{q}) = 1$. However, this perturbation has an approximate interpretation. In fact $\exp(\sqrt{1/2}) = 2.028 \simeq 2$. *A perturbing composition doubling a component, halving another one, and retaining unaltered other components has, approximately, unit e-information*. There are many other e-functions which have unit e-information, but they involve more than two parts. In Figure 2 circles with radii 1, 5, 10 have been plotted. The smallest one is the loci of e-functions with unit e-information.

4. Acquisition of information from an experiment

The fire scenario briefly described in previous sections is studied here in more detail. Consider a suburban zone close to some forest at fire risk. Authorities in charge of safety have to design a mitigation plan for fire affecting the zone. The responsible team may consider several *a priori* hypotheses about the possible states of houses and buildings after the fire. Two of these *a priori* hypotheses have been denoted $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ in Section 2, giving the *a priori* probabilities of a house remaining in the four considered states: no damage (Nod), moderate damage (Mod), severe damage (Sev), collapse or ruin (Col). These two prior distributions of the state of a house correspond to quite different feelings about the effects of the fire. Figure 3 shows $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ as compositional vectors in ilr coordinates, defined in Table 6 by the sign code of an SBP. Observing the prior vectors (circled arrows) in Figure 3, they do not appear as close to orthogonality. Orthogonality of two e-functions means that their information is on unrelated features. In this case, the two considered priors do inform on some common features. Table 1 shows the prior e-functions, their $\mathscr{I}_e$ and the angle they form which is 43^o, thus reflecting the difference in direction of the two priors. The $\mathscr{I}_e$ also differs, since $\mathbf{p}^{(1)}$ is quite more informative than $\mathbf{p}^{(2)}$ (see Table 1).

Table 1: *Two priors of the state of houses in a suburban zone after a fire, with no damage (Nod), moderate damage (Mod), severe damage (Sev) and collapse or ruin (Col) and evidence information $\mathscr{I}_e$.*

e-function	Nod	Mod	Sev	Col	$\mathscr{I}_e$	angle	with
$\mathbf{p}^{(1)}$	0.80	0.15	0.04	0.01	3.24	43.27^o	$\mathbf{p}^{(2)}$
$\mathbf{p}^{(2)}$	0.10	0.10	0.10	0.70	1.69	-43.27^o	$\mathbf{p}^{(1)}$

Next step is studying which inspection procedures are at hand to assess the state of a house after a fire. Two realistic experiments are considered here. The first one consists of a visual inspection of the house by a small trained team. The second is based in airborne photography; the house is identified and its state is assessed on the picture. In what follows, the results of both types of inspection are labelled as the four considered states, adding v (visual) or a (airborne), depending on the type of inspection used. Both types of inspection are uncertain due to several reasons: the inspectors do not know the status of the house previous to the fire; vegetation, burnt or not, can mask relevant details of the structure; access to some parts of the building can be difficult; structural damage can be hidden; there can be errors in the identification of the house, etc. In order to use the result of an inspection to make decisions under a controlled uncertainty, the likelihood of each actual state should be known. Therefore, some assessment of the probability of each outcome of the inspection, conditional to the actual state, is needed. Tables 2 and 3 show the likelihood e-functions (the columns of the tables) for the two types of inspections, visual and airborne, respectively.

Table 2: *Simulated likelihood for the visual inspection of houses. Each column is the likelihood associated with R, i.e. the probabilities of the visual inspection outcome conditional to the actual states,* $\Pr(R|A_i)$. *Row* $\mathscr{I}_e$ *(likelihood) shows the scalar information of the likelihood associated with the observation* R_k. *Rows* $\Pr[R_k^{(i)}]$ *are the probabilities of observing* R_j *given the prior* $\mathbf{p}^{(i)}$, $i = 1,2$, *and the likelihood.*

Actual state	**Visual inspection,** R vNod	vMod	vSev	vCol
No damage (Dam)	0.7665512	0.2001012	0.0333408	0.0000068
Moderate damage (Mod)	0.2000307	0.5999432	0.1201175	0.0799086
Severe damamage (Sev)	0.1176475	0.1765397	0.5293121	0.1765006
Collapse or ruin (Col)	0.0000001	0.1001036	0.1999045	0.6999918
$\mathscr{I}_e$(likelihood)	12.87	1.30	1.99	9.10
$\Pr[R_k^{(1)}]$	0.648	0.258	0.068	0.026
$\Pr[R_k^{(2)}]$	0.108	0.168	0.208	0.516

Table 3: *Simulated likelihood for the airborne inspection of houses. Each column is the likelihood associated with Q, i.e. the probabilities of an outcome of the airborne inspection conditional to the actual states,* $\Pr(Q|A_i)$. *Row* $\mathscr{I}_e$ *(likelihood) shows the scalar information of the likelihood associated to the observation* Q_k. *Rows* $\Pr[Q_k^{(i)}]$ *are the probabilities of observing* Q_j *given the prior* $\mathbf{p}^{(i)}$, $i = 1,2$ *and the likelihood.*

Actual state	**Airborne inspection,** Q aNod	aMod	aSev	aCol
No damage (Nod)	0.6436847	0.3563042	0.0000067	0.0000044
Moderate damage (Mod)	0.3725228	0.5097669	0.0882470	0.0294632
Severe damage (Sev)	0.0860468	0.0967638	0.4408675	0.3763220
Collapse or ruin (Col)	0.0000021	0.0204390	0.2040838	0.7754751
$\mathscr{I}_e$(likelihood)	10.31	2.53	8.99	9.62
$\Pr[Q_k^{(1)}]$	0.574	0.366	0.033	0.027
$\Pr[Q_k^{(2)}]$	0.110	0.111	0.196	0.583

These likelihood tables can be estimated from previous experience in inspection of buildings, which are used as training data for a likelihood model. For instance, a number of houses affected by fire for which the actual state is known were inspected and the result of the inspection was reported. With this kind of data a discriminant analysis of the response of the inspection gives an estimate of the probabilities of the observed state R, conditional to the true state A_i, $\Pr(R|A_i)$. Tables 2 and 3 are the result of a logistic regression on a training set of simulated inspections (not shown in this paper).

In order to represent e-functions in coordinates a contrast matrix (Pawlowsky-Glahn et al., 2015) has been selected. The sign code of the SBP is shown in Table 4. A first look at Tables 2 and 3 reveals the large uncertainty of both types of inspection. Also, some features are clear. For instance, it seems that airborne photography is not efficient in discriminating Nod from Mod and Sev from Col. However, it is able to distinguish

Table 4: *Sign code of SBP defining the coordinates used in the fire scenario.*

coordinate	Nod	Mod	Sev	Col	Expression
1	-1	-1	$+1$	$+1$	$\log(\sqrt{\text{Sev Col}}/\sqrt{\text{Nod Mod}})$
2	-1	$+1$	0	0	$(1/\sqrt{2})\log(\text{Mod}/\text{Nod})$
3	0	0	-1	$+1$	$(1/\sqrt{2})\log(\text{Col}/\text{Sev})$

quite reliably between the two pairs of states. These kinds of interpretation can be improved by computing and representing each likelihood e-function in coordinates, so that the direction and strength of the information are better shown. Figure 2 shows the likelihood e-functions in the ilr-coordinates defined by the SBP coded in Table 4. Although the choice of the SBP is arbitrary and the results of the analysis do not depend on the selected basis, the SBP shown in Table 4 tries to remark the order of damage, from small (-1) to large ($+1$). Two projections are used for the three-dimensional picture: first and second ilr-coordinates (left panel) and first and third ilr-coordinates. Likelihood e-functions are represented by red and blue arrows associated with the visual (v) and airborne (a) inspections, respectively. The length of the arrows are the corresponding scalar information $\mathscr{I}_e$. The first observation is that inspections resulting in no damage (vNod, aNod) or in collapse or ruin (vCol, aCol) are more informative (all of them exceed 5 units of information; see Tables 2 and 3) than the moderate damage outcomes (vMod,aMod). This is due to the fact that Nod and Col observations in both experiments almost exclude the opposite state, Col and Nod, respectively; alternatively vMod, vSev, aMod do not exclude any actual state and they are less resolutive. The most important difference in information between the visual and airborne inspection is related to the severe damage outcome (vSev, aSev). The aSev outcome is relatively much more infor-

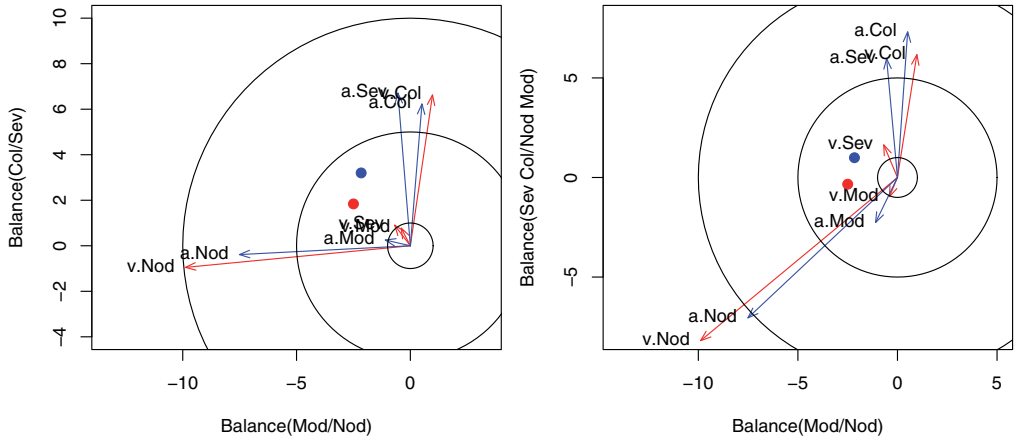


Figure 2: *Likelihood functions as compositions in coordinates. Circles of radius 1, 5, 10. Visual inspection, red arrows; Airborne inspection, blue arrows. Projection first and second coordinates, left panel; first and third coordinates, right panel. Filled markers are the vector averages of the Likelihood functions; red, blue correspond to visual and airborne inspections.*

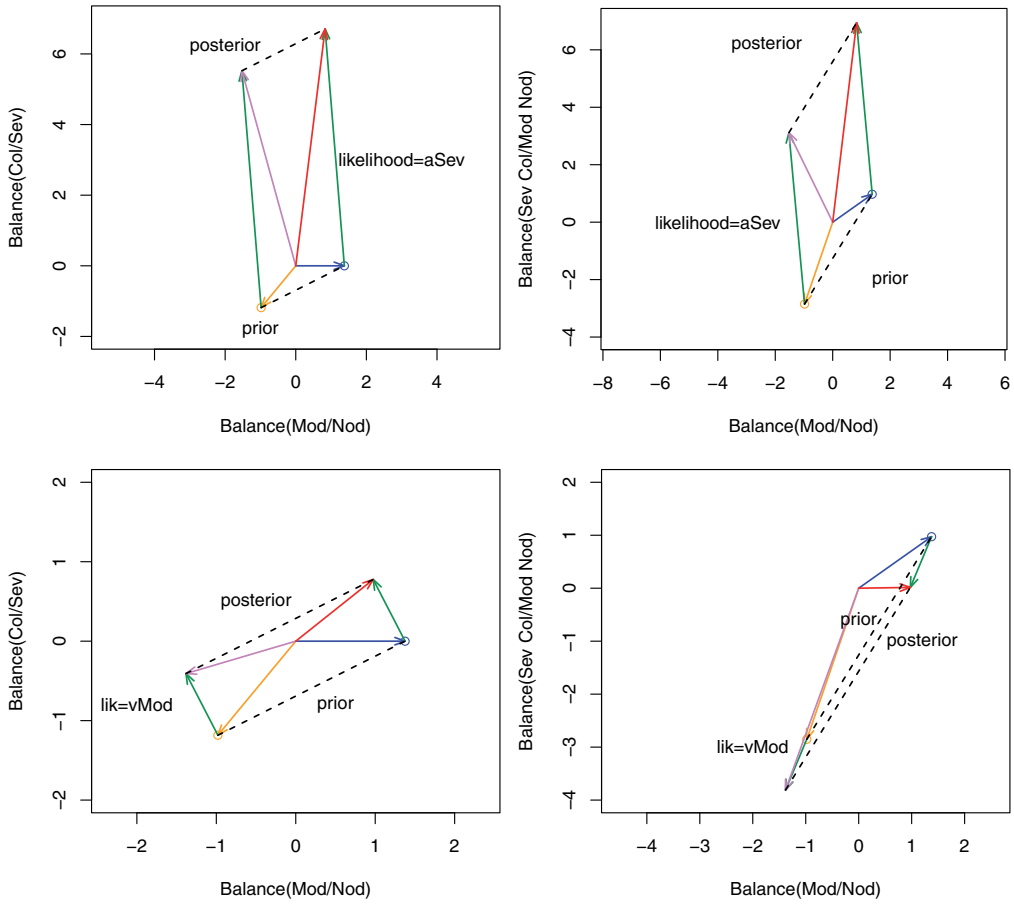


Figure 3: *Bayesian updating: two different priors* $\mathbf{p}^{(1)}$ *(orange, end arrow circled) and* $\mathbf{p}^{(2)}$ *(blue, end arrow circled) are updated with two likelihood cases corresponding to aSev (top panels) and vMod (bottom panels). Left panels show the projection on coordinates 2 and 3 and right panels show projection on coordinates 2 and one as ordered in Table 4. Likelihood (green) is added as a vector to prior. Obtained posteriors* $\mathbf{f}^{(1)}$ *(violet) and* $\mathbf{f}^{(2)}$ *(red) are linked by dotted lines. Priors are also linked by a dotted line to show the parallelogram rule.*

mative, in the scalar sense, than vSev. However, the informative strength of aSev is at the price that aSev gives information that can be confounded with aCol (also with vCol).

The disposition of the likelihood e-functions in both inspections also reveals weaknesses in the design of the inspections. The likelihood arrows in Figure 2 are shifts applied to the prior e-functions. A good design of the experiments should be able to shift the prior in any direction in the three dimensions. Note the inability of these likelihood functions to shift the posterior towards positive values of the balance (Mod/Nod) (second coordinate in Table 4, see Appendix A for further explanation) or negative values of the balance (Col/Sev) (third coordinate in Table 4).

Figure 3 shows the Bayesian updating of the two proposed priors, ($\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$), using outcomes aSev (observed severe damage in the airborne inspection) and vMod (observed moderate damage in the visual inspection) for updating. Top panels of Figure 3 show the two considered priors $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ (Table 1) updated by the likelihood corresponding to the observation of severe damage in the airborne inspection (aSev) in two coordinate projections. The main features are: (a) Likelihood e-functions are not parallel to both priors; consequently, prior assumptions are not confirmed by the observation. It is important to note that parallelism of e-functions would mean that prior assumptions are confirmed by the observations; alternatively, orthogonality of two e-functions means that the e-information they convey do not interact or, more intuitively, they are about different aspects of the scenario. (b) The likelihood is more informative than the two considered priors, i.e. $\mathscr{I}_e(\mathbf{q}) > \mathscr{I}_e(\mathbf{p}^{(k)})$, $k = 1,2$ (See also Tables 1 and 3). (c) The updating hardly modifies the prior coordinate balance of moderate damage (Mod) over no damage (Nod), as the likelihood is almost in the plane defined by the other two coordinates.

Bottom panels of Figure 3 show the two considered priors, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, updated with the likelihood corresponding to the observation of moderate damage in the visual inspection (vMod) in the same projections shown in the top panels. The situation is different from the previous case. Again, the observation does not clearly confirm any of the two priors considered, but the length of the likelihood, $\mathscr{I}_e(\mathbf{q})$, is now smaller than that of the priors: $\mathscr{I}_e(\mathbf{q}) = 1.30$, while $\mathscr{I}_e(\mathbf{p}^{(1)}) = 3.24$, $\mathscr{I}_e(\mathbf{p}^{(2)}) = 1.69$, thus providing a weak change of evidence information from prior to posterior.

Evaluation of visual and airborne inspections

Up to now, only effects of a given observation have been examined. However, decision makers are commonly interested in the evaluation of the available types of inspection, both to know the economical implications of conducting each inspection and how informative they are. Thus, they are interested in the initial question of *which of the two inspections is more informative?* This question was addressed both by Lindley (1956) and in the context of evidence-functions by Egozcue and Pawlowsky-Glahn (2011). In both contributions an average of information provided by possible results of the experiments is proposed. However, it can be discussed which kind of average is more convenient, or which weights are adequate. Here information has a vectorial character, as proposed in Section 2, and accordingly we are primarily concerned with vector averages.

A first possibility is to ignore the probability of each result of an experiment (inspection in our case). This is like considering the experiment outside its context. If the possible likelihood e-functions of the experiments are $\mathbf{q}_k$, $k = 1,2,\ldots,K$ (in the particular case of the considered inspections $K = 4$), the vector average of the likelihood e-functions is

$$\overline{\mathbf{q}} = \frac{1}{K} \odot \bigoplus_{k=1}^{K} \mathbf{q}_k \ ,$$

which is the compositional centre of the set of possible likelihood e-functions. When the e-functions are expressed in coordinates, this is simply the average of the coordinates. These averaged likelihood e-functions are represented in Figure 2 with red and blue markers for the visual and the airborne inspections. If $\overline{\mathbf{q}}$ is not close to the neutral element, it points out that the experiment is quite unable to shift the posterior in the opposite direction. This is the case of both inspections in this example. This motivates the name of *e-information bias* for $\overline{\mathbf{q}}$ or for its norm $\mathscr{I}_e(\overline{\mathbf{q}})$. An experiment with $\overline{\mathbf{q}}$ near the neutral element has the possibility to update the prior e-functions in any direction and is here called *e-information unbiased experiment*.

Common sense points out that the informative value of an experiment depends on the probability of obtaining any outcome. This requires to put the experiment in a particular probabilistic context, which is completely described when the prior e-function is given. In fact, assume that L is a $(K,D)-$matrix with entries $\Pr(R_k|A_i)$, where R_k are the possible outcomes of the experiment R. Tables 2 and 3 show examples of such matrices for the visual and airborne experiments. Matrix multiplication of L and prior probabilities $\mathbf{p}$ give the marginal probabilities for R_k, $\Pr(R_k)$, known as *predictive probabilities* for the observations R_k. The probabilistic weighted average of likelihood e-function is

$$E_R[\mathbf{q}] = \bigoplus_{k=1}^{K} (\Pr(R_k) \odot \mathbf{q}_k) \ , \tag{8}$$

which is the mean likelihood of an experiment in a given probabilistic context. Note that once the prior probabilities and the matrix L are given, the predictive probabilities are also determined. The mean likelihood e-function and its norm, $\mathscr{I}_e(E_R(\mathbf{q}))$, can be considered suitable descriptors of the information provided by an experiment. They can be used to compare experiments.

There are more possibilities of averaging information provided by an experiment. One of them is to average scalar values of $\mathscr{I}_e(\mathbf{q}_k)$. However, a discussion on which is an appropriate scale for $\mathscr{I}_e(\mathbf{q}_k)$ is convenient. In general, the scale of $\mathscr{I}_e(\mathbf{q}_k)$ can be transformed by a monotonous, invertible function $\phi : \mathbb{R}_+ \to \mathbb{R}$ and, then, one can proceed to a weighted average of the transformed values of $\phi(\mathscr{I}_e(\mathbf{q}_k))$. For a general ϕ, it is

$$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})] = \phi^{-1}\left(\sum_{k=1}^{K} \Pr(R_k)\ \phi(\mathscr{I}_e(\mathbf{q}_k)) \right) \ . \tag{9}$$

The scaling function for averaging scalar information has been used by Egozcue and Pawlowsky-Glahn (2011). Table 5 reports some of the available options of scaling functions ϕ. These options are used to evaluate the mean information in Equation (9), provided by the visual and airborne inspections in the fire scenario, and are also reported in Table 5.

Table 5: *Values of mean of scalar e-information $E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$ (9) for the visual and airborne inspections. Probabilities of outcome $\Pr^{(2)}[R_k]$ are reported in Tables 2 and 3.*

Inspection		**Visual**		**Airborne**	
outcome pr.		$\Pr^{(1)}[R_k]$	$\Pr^{(2)}[R_k]$	$\Pr^{(1)}[Q_k]$	$\Pr^{(2)}[Q_k]$
ϕ name	$z=\phi(x)$	$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_Q^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_Q^{\phi}[\mathscr{I}_e(\mathbf{q})]$
identity	$z=x$	2.262	1.681	1.850	2.197
square	$z=x^2$	5.249	3.931	4.139	4.534
neg. exp.	$z=\exp(-x)$	3.914	3.986	4.918	6.107
logarithm	$z=\log x$	1.579	1.493	1.573	1.695
square root	$z=\sqrt{x}$	0.488	0.369	0.425	0.535

Examining these results, one realizes that the mean values depend strongly on the used scaling function, and also on the probabilities of the outcome of the inspection (see values in Tables 2 and 3), which at the same time depend on the prior e-function selected. A second conclusion is that for each scaling function ϕ the most informative inspection depends on the prior. For instance, for ϕ being the identity, and for prior $\mathbf{p}^{(1)}$ and outcome probabilities $\Pr^{(1)}[R_k]$ and $\Pr^{(1)}[Q_k]$, the visual inspection is moderately more informative than the airborne inspection. The situation is reversed for the prior $\mathbf{p}^{(2)}$ and its corresponding outcome probabilities $\Pr^{(2)}[R_k]$ and $\Pr^{(2)}[Q_k]$.

In Table 5 two ϕ options deserve a comment. First, the negative exponential, which considers a monotonous decreasing function. The transformed values $\phi(\mathscr{I}_e(\mathbf{q}_k))$ no longer mean information but a measure of uncertainty or entropy. Accordingly, the average in Equation (9) is a mean value of uncertainties. When transforming back with $\phi^{-1}=-\log$, the mean measure of uncertainty is again translated into e-information. This approach seems quite appealing, but requires further research.

Also, in Table 5, the option $\phi=\log$ may be interesting when a relative scale is assumed for the scalar e-information. However, the relative scale can also be valid for large values of all $\phi(\mathscr{I}_e(\mathbf{q}_k))$ of the experiment. This is due to the fact that the value $\mathscr{I}_e(\mathbf{q}_k)=0$ assigned to the neutral likelihood is attainable, and the relative scale assumed is then nonsensical.

5. Conclusions and further research

The discrete case of Bayesian updating has been considered as a paradigm of information acquisition. Prior information, coded as a probability function, is changed into a final or posterior probability function when the discrete likelihood corresponding to an outcome of an experiment is used in the Bayes' formula. The central idea is that prior, posterior probability functions and, importantly, the discrete likelihood are considered compositions represented in the simplex. The simplex, endowed with the Aitchison geometry, is a Euclidean vector space. The three functions have the characteristics re-

quired by the Aitchison geometry of compositions, thus motivating the common name of evidence functions (e-functions). In this context, Bayes' formula appears exactly as a perturbation of compositions, prior perturbed with likelihood e-functions gives the posterior e-function as a result. The fact that perturbation is the vector sum (group operation) in the Aitchison geometry implies a number of properties; among them, vectors, e-functions in this case, can be represented in (Cartesian) coordinates, thus providing intuitive representations and easy computing of metrics (projections, distances, norms). The conclusion is that information, acquired through Bayes' formula, is a vector magnitude better than a scalar one, as traditionally assumed. Another consequence of this vector approach is that information can be conceived not only for prior and posterior probability functions, but also for likelihood functions which, at the end, is the vector difference between the posterior and the prior.

Generically, vectors have a direction and a modulus or norm. The same is valid for e-functions, which represent a direction of the evidence in the space of compositions and a strength of the evidence, which can be measured as the norm of the e-function. This scalar measure of information may be worth in applications and, accordingly, the norm of e-functions (e-information for short) is taken as a scalar measure of the information conveyed by an e-function. The vectorial character of e-functions introduces some changes in the traditional scalar measures of uncertainty (entropy) or in their counterpart of information. Some intricacies of standard information theory are easily overcome by the Euclidean geometry. For instance, the perturbation-subtraction of e-functions or their distance can advantageously replace divergences or mutual information.

A fire scenario has been used to introduce two kinds of inspection of houses. Questions as simple as *which outcome of the inspection is the most informative* or *which of the two inspections is the most informative?* motivate discussions that require simple operations in the Aitchison geometry. However, different kinds of averages of information provided by the likelihood of an experiment have their own interpretations. The main conclusion is that sensible averages of e-information of an experiment depend on the probabilities of observing the results, which at the same time are determined by the prior probabilities.

The theory and applications of information in evidence functions is not fully developed. A brief description of three possible research directions follows.

The continuous case. The generalization of the log-ratio approach of compositional data to the analysis of density functions, including probability densities, is available (Egozcue, Díaz-Barrero and Pawlowsky-Glahn, 2006, Boogaart, Egozcue and Pawlowsky-Glahn, 2010, Egozcue et al., 2013, Boogaart, Egozcue and Pawlowsky-Glahn, 2014). As in the discrete case, Bayes' theorem consists of the perturbation of the prior density by the likelihood. The continuous e-functions are densities of positive measures, and they are included in infinite dimensional vector spaces called Bayes spaces. Orthogonal projections of e-functions in reduced dimensions are safely introduced when the Bayes space has a Hilbert space structure. In the continuous case, Bayes Hilbert spaces provide

orthonormal coordinates which are Fourier coefficients with respect to bases easily constructed. Some applications have been developed in the framework of geostatistics and functional data (e.g. Menafoglio, Guadagnini and Secchi, 2016, Menafoglio, Grasso, Secchi and Colosimo, 2018), but information applications are still pending.

Weighting e-functions. The theory of Bayes Hilbert spaces (Egozcue et al., 2006, Boogaart et al., 2014) requires a reference (probability) measure of the space. This is specially important when the densities (e-functions) considered have an unbounded support. For interval supported densities and for finite discrete support (compositions) a uniform reference measure is almost automatically adopted. However, this is not the case for infinite supports. This situation suggests that in the interval and compositional cases, adopting a non-uniform reference measure is possible, and in some cases even advisable, thus causing a weighting, in the metrics of the Aitchison geometry, of the information assigned to evidence functions. The way of changing the reference measure for compositions was introduced by Egozcue and Pawlowsky-Glahn (2016), but this approach should be developed and extended to continuous e-functions. In particular, the relationship between prior e-functions and reference measure require further study.

Connections with Dempster-Shafer theory of belief functions. An extensive summary of the theory of belief functions, mainly due to A. P. Dempster and G. Shafer can be found in Yager and Liu (2008), or in the book of Shafer (1976). Belief functions in Dempster theory are operated by Dempster's rule of combination of beliefs (Yager, 1987). Although the support of belief functions is not that of e-functions, the combination of belief functions is just a perturbation, similar to the Bayes' formula in Equation (1). This suggests that belief functions can be viewed as compositions, and the theory here exposed can be extended to belief functions. From this starting point, there is a plea of ideas that deserve attention, like the meaning of orthogonality of e-functions and of belief functions. They seem to be related to exchangeability and independence when using Bayes' formula. These are avenues that should be studied in the future.

Acknowledgements

The authors thank two anonymous referees for their constructive comments which helped improve the manuscript. This work was supported by grants MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund.

A Aitchison geometry

Based on the definitions of perturbation, powering and distance for compositions by Aitchison (1982, 1986), the set of D-part compositions, represented in the simplex $\mathbb{S}^D$,

admits a Euclidean vector space structure (Billheimer et al., 2001, Pawlowsky-Glahn and Egozcue, 2001), which was termed *Aitchison geometry* in the latter reference.

The main elements of this geometry are the vector space operations, perturbation and powering, the metric elements, inner product, distance and norm, and the coordinates for the representation of compositions. In this Appendix A a quick operative reference of these elements is presented. A more comprehensive exposition can be found elsewhere (e.g. Pawlowsky-Glahn et al., 2015, and references therein).

Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_D)$ be D-part compositions represented in $\mathbb{S}^D$. Their perturbation and the powering by a real constant α, are

$$\mathbf{x} \oplus \mathbf{y} = \mathscr{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D) \quad , \quad \alpha \odot \mathbf{x} = \mathscr{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha) \, ,$$

where $\mathscr{C}$ is the closure operation which normalizes the composition to unit sum. With these operations, $\mathbb{S}^D$ is a $(D-1)$-dimensional vector space. Compositions are frequently represented using the *centered log-ratio* (clr) coefficients and *isometric log-ratio* (ilr) coordinates (Egozcue et al., 2003). The clr transformation of $\mathbf{x}$ is

$$\mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \log \frac{x_2}{\mathrm{g_m}(\mathbf{x})}, \ldots \log \frac{x_D}{\mathrm{g_m}(\mathbf{x})} \right) \, ,$$

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the arguments. From the clr coefficients, the composition $\mathbf{x}$ is retrieved by

$$\mathbf{x} = \mathscr{C} \exp(v_1, v_2, \ldots, v_D) \quad , \quad v_i = \mathrm{clr}_i(\mathbf{x}) = \log(x_i / \mathrm{g_m}(\mathbf{x})) \, ,$$

where exp operates componentwise. Note that $\sum_{i=1}^D v_i = 0$.

The ilr coordinates are computed from a $(D, D-1)$ *contrast matrix* V with the properties

$$V^\mathsf{T} V = I_{D-1} \quad , \quad V V^\mathsf{T} = I_D - \frac{1}{D} \mathbf{1}_D \mathbf{1}_D^\mathsf{T} \, , \tag{10}$$

where I_k is the (k,k) identity matrix and $\mathbf{1}_D$ is a column of D unitary entries. Then, the ilr-coordinates associated with V, and with its inverse transformation, are

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \log(V^\mathsf{T} \mathrm{clr}(\mathbf{x})) \quad , \quad \mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{z}) = \mathscr{C}(\exp(V\mathbf{z})) \, ,$$

where $\mathrm{clr}(\mathbf{x}) = \mathbf{v}$ is considered as a column for matrix multiplication. The meaning of these two transformations, clr and ilr, becomes clear after introducing the metric elements of the Aitchison geometry. The Aitchison inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \log \frac{x_i}{x_j} \cdot \log \frac{y_i}{y_j} = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle_e = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}) \rangle_e \, , \tag{11}$$

where $\langle \cdot, \cdot \rangle_e$ denotes the ordinary Euclidean inner product in $\mathbb{R}^D$ when using clr, and in $\mathbb{R}^{D-1}$ when applied to ilr's. From the Aitchison inner product in Equation (11), both the

Aitchison norm, $\|\mathbf{x}\|_a = (\langle \mathbf{x},\mathbf{x}\rangle_a)^{1/2}$, and the Aitchison distance $\mathrm{d}_a(\mathbf{x},\mathbf{y}) = (\|\mathbf{x}\ominus\mathbf{y}\|_a)^{1/2}$ are readily obtained. Some useful expressions for the squared Aitchison distance are

$$\mathrm{d}_a^2(\mathbf{x},\mathbf{y}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)^2 = \sum_{i=1}^{D}[\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\mathbf{y})]^2$$

$$= \sum_{i=1}^{D-1}[\mathrm{ilr}_i(\mathbf{x}) - \mathrm{ilr}_i(\mathbf{y})]^2 \,, \tag{12}$$

and for the squared Aitchison norm

$$\|\mathbf{x}\|_a^2 = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log\frac{x_i}{x_j}\right)^2 = \sum_{i=1}^{D}[\mathrm{clr}_i(\mathbf{x})]^2 = \sum_{i=1}^{D-1}[\mathrm{ilr}_i(\mathbf{x})]^2 \,, \tag{13}$$

where $\mathrm{clr}_i(\mathbf{x})$ and $\mathrm{ilr}_i(\mathbf{x})$ denote the components of $\mathrm{clr}(\mathbf{x})$ and $\mathrm{ilr}(\mathbf{x})$ respectively.

From these definitions, it is clear that V contains the clr coefficients of the compositions of the selected basis in $\mathbb{S}^D$. Then, the condition $V^{\mathrm{T}}V = I_{D-1}$ implies the orthonormality of the basis and, consequently, the corresponding ilr-coordinates are Cartesian coordinates representing the composition. Both clr and ilr define isometries from $\mathbb{S}^D$ onto $\mathbb{R}_0^D$ (real D-vectors which components add to zero) and $\mathbb{R}^{D-1}$, respectively. This can be summarized as

$$\mathrm{clr}(\alpha\odot\mathbf{x}\oplus\mathbf{y}) = \alpha\cdot\mathrm{clr}(\mathbf{x}) + \mathrm{clr}(\mathbf{y}) \quad , \quad \mathrm{ilr}(\alpha\odot\mathbf{x}\oplus\mathbf{y}) = \alpha\cdot\mathrm{ilr}(\mathbf{x}) + \mathrm{ilr}(\mathbf{y}) \,,$$

and

$$\langle \mathbf{x},\mathbf{y}\rangle_a = \langle \mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})\rangle_e = \langle \mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y})\rangle_e \,,$$

$$\mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}_e(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})) = \mathrm{d}_e(\mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y})) \;\; , \;\; \|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\|_e = \|\mathrm{ilr}(\mathbf{x})\|_e \,,$$

where subscripts a mean Aitchison geometry, and subscripts e mean ordinary Euclidean geometry. Note that real operations involving clr coefficients are carried out in $\mathbb{R}^D$, while those involving ilr correspond to $\mathbb{R}^{D-1}$.

A practical way of constructing ilr-coordinates, i.e. of obtaining the contrast matrix V, is using sequential binary partitions (SBP) of the compositions. This technique (Egozcue and Pawlowsky-Glahn, 2005) consists of separating into two (non overlapping) groups the parts of a composition, for instance, marking the parts in each group with a $+1$ and with a -1 otherwise. The partition is repeated in each group generated in previous steps. A typical way of coding the SBP is shown as example in Table 6.

The sign code of the SBP is given in the $(D, D-1)$ matrix $\Theta = [\theta_{ij}]$, where the code component ij corresponds to the sign of x_i in the j-th ilr coordinate. Each step of partition corresponds to an element $\mathbf{e}_j$ of the orthonormal basis, and the corresponding j-th ilr-coordinate is computed as

$$b_j = \mathrm{ilr}_j(\mathbf{x}) = \sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \log \frac{(\prod_{\theta_{ij}=+1} x_{ij})^{1/n_+}}{(\prod_{\theta_{ij}=-1} x_{ij})^{1/n_-}} \;,\; j = 1,2,\ldots,D-1 \qquad (14)$$

where n_+ and n_- are the number of plus signs and minus signs, respectively. Note that the expression $(\prod_{\theta_{ij}=+1} x_{ij})^{1/n_+}$ in the numerator of the fraction in Equation (14) is the geometric mean of the elements x_{ij} which are marked with a $+1$ in the j-th partition. Similarly the expression in the denominator for elements marked with a -1. The coordinates b_j have a particularly simple form: they are proportional to log-ratios of geometric means of groups. Due to this fact, they are called *balances* between the corresponding groups of parts (Egozcue et al., 2003, Egozcue and Pawlowsky-Glahn, 2005). An abbreviated way of denoting balances is to enumerate the parts in the numerator and denominator separated by a slash. For instance, the $j = 2$ balance coded as in Table 6 would be denoted as balance$(x_2, x_D / x_3, x_4, \ldots, x_{D-1})$. The elements of the contrast matrix, v_{ij} are null if $\theta_{ij} = 0$ and, for $\theta_{ij} = +1$ and $\theta_{ij} = -1$,

$$v_{ij} = \frac{\theta_{ij}}{n_+}\sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \quad , \quad v_{ij} = \frac{\theta_{ij}}{n_-}\sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \;,$$

respectively. Note that, if $\mathbf{e}_j$ is the j-th element of the basis, then $\mathrm{clr}(\mathbf{e}_j) = (v_{1j}, v_{2j}, \ldots, v_{Dj})^\mathsf{T}$.

Table 6: *Sign code for a SBP of a D part composition to compute coordinates* $\mathrm{ilr}_j(\mathbf{x})$. *As an example, first partition separates* x_1 $(+1)$ *from the rest of parts; the second step separates* x_2 *and* x_D *from parts previously marked with* -1*; parts not participating in this partition step are labelled as* 0. *Take the* $+1$, -1, 0 *codes as entries of a matrix* Θ^T.

	sign code matrix Θ^T						
j	x_1	x_2	x_3	x_4	…	x_{D-1}	x_D
1	$+1$	-1	-1	-1	…	-1	-1
2	0	$+1$	-1	-1	…	-1	$+1$
3	0	$+1$	0	0	…	0	-1
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
$D-1$	0	0	$+1$	-1	…	0	0

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44, 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70, 57–65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London (UK): Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).

Aitchison, J. (1994). *Multivariate analysis and its applications*, Volume 24 of *Lecture Notes–Monograph Series*, Chapter Principles of compositional data analysis, pp. 73–81. Hayward, CA: Institute of Mathematical Statistics.

Aitchison, J. and Kay, J. (1975). Principles, practice and performance in decision-making in clinical medicine. In K. C. Bowen and D. G. White (Eds.), *Proceedings of the 1973 NATO conference on The Role and Effectiveness of Decision Theories in Practice*, London (GB). English Universities Press.

Aitchison, J., Kay, J. W. and Lauder, I. J. (2005). *Statistical Concepts and Applications in Clinical Medicine*. Chapman and Hall/CRC.

Ash, R. B. (1990). *Information theory*. Dover, New York; first puiblished by J. Wiley & Sons, 1965.

Barceló-Vidal, C. and Martín-Fernández, J.-A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45, 57–71.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology*, CD-ROM.

Benjamin, J. R. and Cornell, C. A. (1960). *Probability, Statistics and Decision for Civil Engineers*. McGraw Hill Companies.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96, 1205–1214.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57, 269–326.

Boogaart, K. G. v., Egozcue, J. J. and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT - Statistics and Operations Research Transactions*, 34, 201–222.

Boogaart, K. G. v., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian and New Zealand Journal of Statistics*, 56, 171–194.

Egozcue, J. J., Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica (English Series)*, 22, 1175–1182. DOI: 10.1007/s10114-005-0678-2.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*,37, 795–828.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2011). Evidence information in bayesian updating. In J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego (Eds.), *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011)*. CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2016). Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, 45, 25–44.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2018). *Modelling compositional data. The sample space approach*. Chapter 4 in Fifty years if IAMG, D. Sagar, Q. M. Chen and F. Agterberg (Eds.), Springer.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *TEST*. submitted.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Egozcue, J. J., Pawlowsky-Glahn, V., Templ, M. and Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics –Theory and Methods*, 44, 3978–3996.

Egozcue, J. J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M. I. and van den Boogaart, K. G. (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas (RACSAM)*, 107, 475–486. DOI 10.1007/s13398-012-0082-6.

Gray, R. M. (2011). *Entropy and Information Theory* (2nd ed.). Springer, New York.

Hartley, R. V. L. (1928). Transmission of information. *Bell Systems Technical Journal*, 7, 535–563.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York, NY (USA).

Kullback, S. (1997). *Information Theory and Statistics, an unabridged republication of the Dover 1968 edition*. Dover publications, Minetola.

Kullback, S. and Leibler, R. A. (1951a). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.

Kullback, S. and Leibler, R. A. (1951b). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27, 986–1005.

Martín-Fernández, J. A. (2001). *Medidas de diferencia y clasificación no paramétrica de datos composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E).

McMillan, B. (1953). The basic theorems of information theory. *Annals of Mathematical Statistics*, 24, 196–219.

Menafoglio, A., Grasso, M., Secchi, P. and Colosimo, B. M. (2018). Profile monitoring of probability density functions via simplicial functional PCA with application to image data. *Technometrics* online February 12, 2018.

Menafoglio, A., Guadagnini, A. and Secchi, P. (2016). Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. *Water Resources Research*, 52, 5708–5726.

Pawlowsky-Glahn, V. and Buccianti, A. (Eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15, 384–398.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Statistics in practice. John Wiley & Sons, Chichester UK.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton NJ, USA.

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379–423, 623–656.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press. Urbana.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41, 93–137.

Yager, R. R. and Liu, L. (Eds.) (2008). *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer, Berlin.

SORT 42 (2) July-December 2018, 125-158 DOI: 10.2436/20.8080.02.72

A contingency table approach based on nearest neighbour relations for testing self and mixed correspondence

Elvan Ceyhan*

Abstract

Nearest neighbour methods are employed for drawing inferences about spatial patterns of points from two or more classes. We introduce a new pattern called correspondence which is motivated by (spatial) niche/habitat specificity and segregation, and define an associated contingency table called a correspondence contingency table, and examine the relation of correspondence with the motivating patterns (namely, segregation and niche specificity). We propose tests based on the correspondence contingency table for testing self and mixed correspondence and determine the appropriate null hypotheses and the underlying conditions appropriate for these tests. We compare finite sample performance of the tests in terms of empirical size and power by extensive Monte Carlo simulations and illustrate the methods on two artificial data sets and one real-life ecological data set.

MSC: 62M30, 62G10, 62H11, 62P12, 62H30

Keywords: Association, complete spatial randomness, habitat/niche specificity, independence, random labelling, segregation

1. Introduction

The spatial point patterns in natural populations (in $\mathbb{R}^2$ and $\mathbb{R}^3$) have received considerable attention in statistical literature. Among the frequently studied spatial patterns between multiple classes/species are segregation and association (Dixon, 2002a), and niche specificity pattern (Primack, 1998). Pielou (1961) proposed various tests based on nearest neighbour (NN) relations in a two-class setting, namely, tests of segregation, symmetry, and niche specificity, and also a coefficient of segregation. Inspired by niche specificity and segregation, we introduce new multi-class patterns called *self* and *mixed correspondence* in the NN structure. We use the NN relationships for testing these patterns. In this article, we only use the first NN (i.e., 1-NN) of any point so NN always

* Department of Statistics, North Carolina State University, Raleigh, NC, 27695. e-mail: eceyhan@ncsu.edu
Received: July 2015
Accepted: June 2018

refers to the 1-NN. Furthermore, the terms "class" and "species" are used interchangeably and refer to any characteristic of the subjects such as gender, age group, health condition, etc.

We also propose tests for the spatial patterns of *self* and *mixed correspondence* in the NN structure. These tests are based on a contingency table called a *correspondence contingency table* (CCT) which is constructed using the NN relations in the data. A *base-NN pair* (or simply the *NN pair*) is the pair of points (p_1, p_2) in which p_2 is a NN of p_1, and p_1 is called the *base point* and p_2 is called the *NN point*. The NN pair (p_1, p_2) is called a *self pair*, if both p_1 and p_2 are from the same class, while it is called a *mixed pair*, if p_1 and p_2 are from different classes. *Self correspondence* in the NN structure occurs when there is a tendency for points from a class to be NNs to points from the same class. That is, self correspondence occurs when self NN pairs are more abundant than expected. On the other hand, *mixed correspondence* occurs when there is a tendency for points and their NNs to be from different classes, i.e., mixed NN pairs are more abundant than expected.

There are many methods available for testing various types of spatial patterns in literature. These spatial tests include Pielou's test of segregation (Pielou, 1961), Ripley's K-function (Ripley, 2004), or J-function (van Lieshout and Baddeley, 1999), and so on. Some of these methods are based on nearest neighbour (NN) relations between the points in the data set (Dixon, 2002b). For example, Clark and Evans (1954) use the mean distance of points to their NNs in a spatial data set to measure the deviations of plant species from spatial randomness and compare the deviations for multiple species. However, in this article, we base our analysis on the class labels of NN pairs as was done in Pielou (1961) and Dixon (1994). An extensive survey for the tests of spatial point patterns is provided by Kulldorff (2006) who categorized and compared more than 100 such tests. These tests are for testing spatial clustering in a one-class setting or testing segregation of points in a multi-class setting. The null hypothesis is some type of spatial randomness and is usually fully specified, but the alternatives are often not so definite, in the sense that for most tests the alternatives are presented as deviations from the null case are of interest as in pure significance tests of Cox and Hinkley (1974); only a few tests specify an explicit alternative clustering scheme. Most of the tests for multiple classes deal with presence or lack of spatial interaction usually in the form of spatial segregation or association between the classes. However, none of the numerous tests surveyed by Kulldorff (2006) are designed for testing correspondence; and the pattern of correspondence and the associated tests are introduced in this article. The tests for assessing the self and mixed correspondence in the NN structure are based on the CCT which can also be constructed by collapsing the nearest neighbour contingency table (NNCT). See Ceyhan (2010, 2008a) for an extensive treatment of NNCT and tests based on it.

We provide the description of correspondence and related patterns, the list of notations and two motivating (artificial) examples in Section 2. A list of abbreviations used in the article is provided in Table 1. We propose the pattern of correspondence together

with the associated tests and the contingency table (i.e., CCT) and the benchmark and the null patterns for correspondence in Section 3 where the asymptotic distributions of the cell counts in the CCT and of the tests based on them are also derived. We prove consistency of the tests in Section 3.3, and provide an extensive empirical size and power analysis by Monte Carlo simulations in Section 4. We also illustrate the methodology on one ecological data set in Section 5 and provide some discussion and guidelines in Section 6.

Table 1: *A list of abbreviations used in the article.*

CSR:	Complete Spatial Randomness
NN:	Nearest Neighbour
NNCT:	Nearest Neighbour Contingency Table
RL:	Random Labelling
CCT:	Correspondence Contingency Table

2. Preliminaries

2.1. Spatial Correspondence and Related Patterns

We first introduce the motivating patterns of niche specificity and segregation and then discuss their connection with correspondence.

Niche/habitat specificity is the collection of biotic and abiotic conditions favouring the development, hence existence and abundance of a species on a spatial scale (Ranker and Haufler, 2008). That is, niche specificity is the dependence of an organism on an environment (i.e., niche or habitat). In literature, niche/habitat specificity is also discussed within the context of species diversity under the title of *habitat association* of two or more species (Primack, 1998). Niche specificity is a broad concept and is determined by partitioning of the niche space. Furthermore, niche space has non-spatial coordinates amenable for niche partitioning; e.g., Fargione and Tilman (2005) uses different phenologies resulting in temporal partitioning of the niche space and Werner and Gilliam (1984) incorporate ontogenetic changes (i.e., changes as an individual develops in size) to partition the niche space. However, in this article, we are mainly concerned with the spatial aspect of multi-class interaction patterns.

In a multi-species setting, *segregation* of a species is the pattern in which members of a species occur near members of the same species (Dixon, 1994). Conversely, *association* of a species to another is the pattern in which members of the former species tend to occur near the members of the latter. That is, under segregation, the members of a class or species enjoy the company of the conspecifics, hence form one class clumps or clusters, while under association they tend to coexist with members of other class(es) and form mixed clumps or clusters (see, e.g., Ceyhan, 2008a for more detail).

Niche specificity can be viewed as a factor that accounts for segregation which can account for self correspondence. In a multi-species setting, if each species were confined to its own support/niche, we would expect one-species clumps (which would tend to exclude other species). So if (spatial) niche specificity is in effect for all species in the study region, self correspondence would occur (i.e., self NN pairs would be more abundant than mixed pairs). On the other hand, if niche specificity is in effect for one species, then that species would exhibit segregation from the rest of the species. Self correspondence is much closer to the concept of segregation compared to niche specificity, as self correspondence and segregation are both based on the spatial proximity of the conspecifics. Self correspondence in the NN structure pertains to the NN pair types as self or mixed for each class among all base-NN pairs and thus to a supra-species characteristic. However segregation is a pattern at the species level, in the sense that one can only talk of segregation of a species from another or others. That is, in a multi-class or multi-species setting, self correspondence refers to the NN preference of species for all species combined and so it is intended to measure whether species prefer their conspecifics in a cumulative fashion, i.e., for all species taken into account together. Thus, segregation is defined at species level, while self correspondence is defined at multi-species level; and the two patterns are related but different in the sense that, e.g., all species together might exhibit self correspondence without significant segregation for any of the species. But segregation of all or most species will usually substantiate the presence of self correspondence, hence segregation can be viewed as a factor that accounts for self correspondence. Lack of segregation might indicate mixed correspondence, which may or may not imply association, since for association one needs to consider each pair of species separately and test the interaction between the two species in the pair. Lack of segregation is guaranteed to imply presence of association in the two-class setting only.

2.2. Notation

For convenience to the reader, following the example of Vichi and Saporta (2009), we provide the notation and terminology used in the article below.

X and Y	class labels (interchangeably 1 and 2, respectively);
$\mathscr{X}_n$ and $\mathscr{Y}_m$	a data set of size n from class X and a data set of size m from class Y;
$\mathscr{W}_n$	represents the combined data set for the CSR setting, and the background points for the RL setting;
$\mathscr{D}_n$	the set of ordered pairs (W_i, L_i), where W_i stands for the location of the point and L_i stands for the corresponding class label;
S_i	the number of self base-NN pairs for class i;
M_i	the number of mixed base-NN pairs with base point being from class i;
S and M	sum of the first column (for self pairs) i.e. $S = \sum_{i=1}^{k} S_i$ and sum of the second column (for mixed pairs) i.e. $M = \sum_{i=1}^{k} M_i$ in the CCT;

N_{ij}	the observed frequency of category (i, j) in the NNCT, i.e., the number of (base,NN) pairs in which base class is i and NN class is j;
C_i	sum of column j in the NNCT;
R and Q	twice the number of reflexive pairs and the number of points with shared NNs, which occurs when two or more points share a NN;
Q_l	the number of points that serve as a NN to other points l times;
Z_{S_i} and Z_S	the test statistics for cell $(i,1)$ in the CCT and for sum of the self column, S;
Z_{ii}	the cell-specific tests for cell (i,i) in the NNCT analysis;
$\mathbf{S}$	the vector of combined S_i values (i.e., the self column in the CCT), i.e., $(S_1, S_2, \ldots, S_k)$;
$\mathbf{\Sigma_S}$	the variance-covariance matrix of $\mathbf{S}$;
$\mathscr{X}_C$	the (quadratic form) test statistic for the correspondence;
$\mathscr{X}_D$	the overall segregation test due to Dixon;
$\mathbf{N}$ and $\mathbf{\Sigma_N}$	the vector of entries of NNCT concatenated row-wise and its covariance matrix;
A^-	the generalized inverse of a matrix A;
$\chi^2_{\nu,\alpha}$	the $100\alpha^{th}$ percentile of χ^2 distribution with ν degrees of freedom;
N_{mc}	the number of Monte Carlo samples generated for the empirical size and power comparison of the tests;
$\widehat{\alpha}_T$	the empirical size estimate of a test statistic, T, at level $\alpha = 0.05$;
$\widehat{\alpha}_{T_1,T_2}$	the proportion of agreement in rejecting the null hypothesis between test statistics T_1 and T_2;
$\mathscr{U}(A)$	the uniform distribution on region A;
MatClust(κ, r, μ)	Matérn cluster process with Poisson parameters κ and μ and radius r;
$\widehat{\beta}_T$	the empirical power estimate of a test statistic, T, at level $\alpha = 0.05$;
p_{asy}	the p-value based on the asymptotic approximation (i.e., asymptotic critical value);
p_{rand}	the p-value based on Monte Carlo randomization of the labels on the given locations;
p_{mc}	the p-value based on 10000 Monte Carlo replication of the CSR independence pattern in the study region

2.3. Motivating Examples

To motivate the patterns of self/mixed correspondence and how they can be different from segregation/association, we use two artificial data sets, each of which has three classes (representing tree species) say, X, Y and Z in a square study region. We could also choose examples with two classes, but with two classes only one of the newly intro-

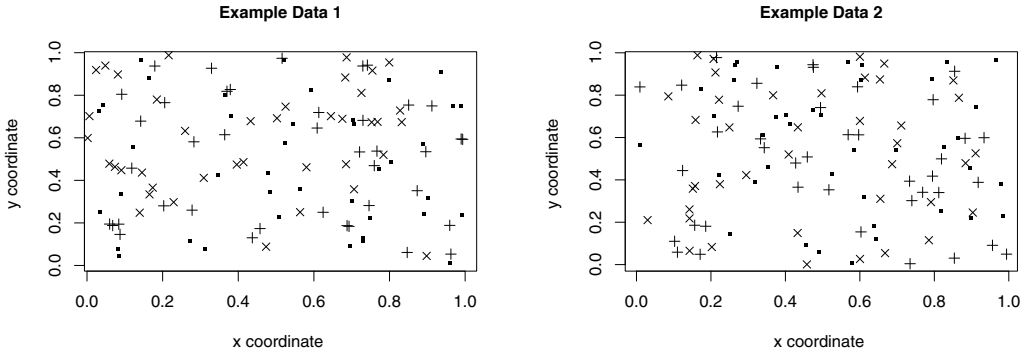


Figure 1: *The scatterplots of the locations of three classes (representing three tree species) in our artificial data set 1 (left) and artificial data set 2 (right). There are 40 points in each class/species.*

duced tests provide new information compared to the existing segregation tests and there are more possibilities of different types of pairwise interactions between classes with three or more classes. Hence it would be more informative to discuss the differences between the tests and patterns of self/mixed correspondence and segregation/association with three or more classes. We generate 40 points for each class and the locations of the points are plotted in Figure 1. The scatter plot for artificial data set 1 on the left is suggestive of mild self correspondence and segregation with number of self NN pairs being 54 and number of mixed NN pairs being 66. The scatter plot for artificial data set 2 on the right is suggestive of mild mixed correspondence and a lack of segregation with number of self NN pairs being 34 and number of mixed NN pairs being 86. However, these claims are not assessed rigorously yet to attach any significance (or lack of it) to them. We will illustrate the correspondence and segregation patterns and the associated tests using these examples in the following sections.

3. Correspondence in the NN Structure and the Associated Contingency Table

3.1. Benchmark and Null Patterns for Multivariate Spatial Interaction

In this article, we are concerned with the (spatial) interaction between two or more classes of points, particularly with correspondence. For multivariate spatial data analysis, the benchmark pattern is usually complete spatial randomness (CSR) independence or random labelling (RL) (Diggle, 2003) depending on the context. The distinction between CSR independence and RL could be very important in practice. Under CSR independence the (locations of the) points from two classes are *a priori* the result of different processes (for instance, individuals of different species or age cohorts). On the other hand, under RL, some processes affect the individuals of a single population *a posteriori* (for instance, diseased versus non-diseased individuals of a single plant

species) (Goreaud and Pélissier, 2003). Under CSR independence, the points from each class are independently uniformly distributed in the region of interest conditioned on the class sizes. That is, the points from each class are independent realizations from a Homogeneous Poisson Process (HPP) with fixed class sizes (i.e., they are independent realizations from a binomial process). On the other hand, under RL, class labels are independently and randomly assigned to a set of given locations which could be a realization from any pattern such as HPP or some clustered or regular pattern.

For simplicity, we describe the benchmark patterns for the two-class case. Extension to multi-class case is straightforward. In a two-class setting, we label the classes as X and Y (or interchangeably 1 and 2, respectively). Let $\mathscr{X}_{n_1}$ be a data set of size n_1 from class X and $\mathscr{Y}_{n_2}$ be a data set of size n_2 from class Y. Then under CSR independence, we have $\mathscr{X}_{n_1} = \{X_1, X_2, \ldots, X_{n_1}\}$ and $\mathscr{Y}_{n_2} = \{Y_1, Y_2, \ldots, Y_{n_2}\}$ which are independent and are both random samples from $\mathscr{U}(S)$, the uniform distribution on the common support $S \subset \mathbb{R}^d$ for classes X and Y where $\mathbb{R}^d$ is the d-dimensional Euclidean space. Unless stated otherwise, for simplicity and practical purposes, we take $d = 2$ (i.e., consider planar data) throughout the article. We combine $\mathscr{X}_{n_1}$ and $\mathscr{Y}_{n_2}$ into one data set $\mathscr{W}_n = \mathscr{X}_{n_1} \cup \mathscr{Y}_{n_2} = \{W_1, W_2, \ldots, W_n\}$ where $n = n_1 + n_2$. In fact, we consider labeled data points as $\mathscr{D}_n = \{(W_i, L_i) \text{ for } i = 1, 2, \ldots, n\}$ where $L_i \in \{0, 1\}$ or $\{X, Y\}$ are the class labels. Notice that under CSR independence, the randomness is in the locations of the points W_i and the class label is a fixed deterministic characteristic of the point. Under the RL pattern, the class labels or marks are assigned randomly to points whose locations are given. The spatial pattern generating these point locations is referred to as the *background pattern* henceforth. Then $\mathscr{W}_n$ is the given set of locations for n points from the background pattern. We have the pair of observations (W_i, L_i) where $L_i \in \{1, 2\}$ or $\{X, Y\}$ is the class label of the point W_i for $i = 1, 2, \ldots, n$. Then n_1 (resp. n_2) of these W_i points are assigned as class X (resp. class Y) randomly; i.e., the labels L_i are 1 or X approximately with probability n_1/n (resp. 2 or Y with probability n_2/n) independently for $i = 1, 2, \ldots, n$. Under RL, the locations of the points are fixed but the randomness is in the label, L_i, associated with these points.

There are two major types of interaction pattern types as deviation from these benchmark patterns in the multivariate spatial pattern analysis. These interaction patterns are segregation and association. Segregation/association, niche specificity and correspondence are related but different concepts (see Section 2.1), and hence the corresponding null hypotheses are different. Niche specificity might account for or explain the self correspondence or segregation patterns. In particular, if niche specificity occurs at significant levels for each species, then there will be significant segregation for each species and significant self correspondence for all species combined. But if niche specificity occurs for some species (but for other species niche specificity is not significant or these other species exist in mixed groups/clumps or scattered around the study region haphazardly), segregation is operating for these species, while self correspondence may or may not be in effect (e.g., segregation of species may not be strong enough to render self pairs significantly larger than expected, or the associated pairs might hinder the occur-

rence of self correspondence). Hence self correspondence and segregation are different patterns with substantial overlap, but one is not the subset of the other. We provide the explicit forms of the corresponding null hypotheses in the subsequent sections.

3.2. Tests of Correspondence and Their Relation to Segregation

The null case for self or mixed correspondence is that the entries for self (or mixed) pair types in the CCT are as expected under RL or CSR independence.

For a species to exhibit self (resp. mixed) correspondence in the NN structure, self (resp. mixed) NN pairs would be more abundant than expected under RL. To detect such type of pattern, we construct a contingency table where NN pairs are classified as self or mixed for each class. Let S_i be the number of self NN pairs for class i, and M_i be the number of mixed NN pairs with base point being from class i. For simplicity, we assume there are no ties in the NN relations, which occurs with probability one, if $\mathscr{W}_n$ is a random sample from a continuous distribution. Then

$$S_i = \sum_{j\neq i, j=1}^{n} \sum_{i=1}^{n} \mathbf{I}(Z_j \text{ is a NN of } Z_i)\mathbf{I}(L_i = L_j),$$

and

$$M_i = n_i - S_i = \sum_{j\neq i, j=1}^{n} \sum_{i=1}^{n} \mathbf{I}(Z_j \text{ is a NN of } Z_i)\mathbf{I}(L_i \neq L_j).$$

Then the resulting contingency table is a $k \times 2$ contingency table for k classes with first column (called self column) comprising of S_i and the second column (called mixed column) comprising of M_i values. See also Table 3 (left). Notice that row sums are class sizes (i.e., sum of row i is n_i), and sum of the self column is $S = \sum_{i=1}^{k} S_i$ and sum of the mixed column is $M = \sum_{i=1}^{k} M_i$.

Remark 3.1. Ties in the NN Structure. If there are ties in the NN structure, which can happen, e.g., due to truncation of the coordinates of the observations when recording, we can adjust the above formulas for S_i and M_i by inserting a weight term for ties. For instance, we can write $\frac{1}{N_i^{nn}}\mathbf{I}(Z_j \text{ is a NN of } Z_i)$ to account for the ties where N_i^{nn} is the number of NNs of point Z_i. Note that $N_i^{nn} = 1$ with probability 1 when $\mathscr{W}_n$ is a random sample from a continuous distribution. □

The $k \times 2$ CCT is closely related to the $k \times k$ nearest neighbour contingency table (NNCT) based on the same data. Here we provide a brief description of NNCTs (for more detail, see, e.g., Ceyhan, 2008a). NNCTs are constructed using the NN frequencies of classes. Let n_i be the number of points from class i (assumed fixed) for $i \in \{1, 2, \ldots, k\}$ and $n = \sum_{i=1}^{k} n_i$. If we record the class of each point and its NN, the NN relationships fall into the following k^2 categories:

$$(1,1),(1,2),\ldots,(1,k);(2,1),(2,2),\ldots,(2,k);\ldots,(k,k)$$

where in category or cell (i,j), class i is called the *base class*, and class j is called the *NN class*. Denoting N_{ij} as the observed frequency of category (i,j) for $i,j \in \{1,2,\ldots,k\}$, we obtain the NNCT in Table 3 (right). Then,

$$N_{ij} = \sum_{j' \neq i', j'=1}^{n} \sum_{i'=1}^{n} \mathbf{I}(Z_{j'} \text{ is a NN of } Z_{i'}) \mathbf{I}(L_{i'} = i) \mathbf{I}(L_{j'} = j).$$

The number of self pairs for class i is same as the number of base-NN pairs with both base and NN classes are from class i. Hence $S_i = N_{ii}$ and $M_i = n_i - N_{ii}$.

Table 3: *The CCT (left) and the NNCT (right) for k classes.*

		pair type		
		self	mixed	total
base class	class 1	S_1	M_1	n_1
	class 2	S_2	M_2	n_2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	class k	S_k	M_k	n_k
	total	S	M	n

		NN class			
		class 1	$\ldots$	class k	total
base class	class 1	N_{11}	$\ldots$	N_{1k}	n_1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	class k	N_{k1}	$\ldots$	N_{kk}	n_k
	total	C_1	$\ldots$	C_k	n

Table 4: *The CCT (left) and the NNCT (right) for the three classes in our artificial data set 1.*

		pair type		
		self	mixed	total
base class	class 1	18	22	40
	class 2	18	22	40
	class 3	18	22	40
	total	54	66	120

		NN class			
		class 1	class 2	class 3	total
base class	class 1	18	13	9	40
	class 2	11	18	11	40
	class 3	8	14	18	40
	total	37	45	38	120

Table 5: *The CCT (left) and the NNCT (right) for the three classes in our artificial data set 2.*

		pair type		
		self	mixed	total
base class	class 1	7	33	40
	class 2	19	21	40
	class 3	8	32	40
	total	34	86	120

		NN class			
		class 1	class 2	class 3	total
base class	class 1	7	15	18	40
	class 2	10	19	11	40
	class 3	18	14	8	40
	total	35	48	37	120

We present the CCTs and NNCTs for the artificial data sets 1 and 2 in Tables 4 and 5, respectively. For artificial data set 1, the CCT suggests presence of self correspondence with self column entries being higher than expected. Equivalently, the NNCT diagonal entries are higher than expected suggesting presence of segregation of the classes. For artificial data set 2, based on the CCT, we observe that there seems to be mixed cor-

respondence (with NN pairs in which base points are from classes 1 or 2). Likewise, NNCT suggests that classes 1 and 3 are associated with each other, and there is a lack of segregation for these classes, and class 2 points seem to be segregated from points of other classes.

Under RL, we can determine the exact expected values, variances, and asymptotic distributions of the cell counts in the CCT. In particular,

$$\mathrm{E}[S_i] = \mathrm{E}[N_{ii}] = n_i(n_i-1)/(n-1) \text{ and } \mathrm{E}[M_i] = \mathrm{E}[n_i - N_{ii}] = n_i(n-n_i)/(n-1). \quad (1)$$

Furthermore,

$$\mathrm{Var}[S_i] = \mathrm{Var}[N_{ii}] = (n+R)p_{ii} + (2n-2R+Q)p_{iii} + (n^2-3n-Q+R)p_{iiii} - n^2p_{ii}^2 \quad (2)$$

and since n_i are fixed

$$\mathrm{Var}[M_i] = \mathrm{Var}[n_i - N_{ii}] = \mathrm{Var}[N_{ii}] = \mathrm{Var}[S_i].$$

In Equation (2), p_{xx}, p_{xxx}, and p_{xxxx} are the probabilities that a randomly picked pair, triplet, or quartet of points, respectively, are the indicated classes and are given by

$$p_{ii} = \frac{n_i\,(n_i-1)}{n\,(n-1)},\ p_{iii} = \frac{n_i\,(n_i-1)\,(n_i-2)}{n\,(n-1)\,(n-2)},\ p_{iiii} = \frac{n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)}{n\,(n-1)\,(n-2)\,(n-3)}, \quad (3)$$

and R is twice the number of reflexive pairs and Q is the number of points with shared NNs, which occurs when two or more points share a NN. Then $Q = 2\,(Q_2 + 3\,Q_3 + 6\,Q_4 + 10\,Q_5 + 15\,Q_6)$ where Q_l is the number of points that serve as a NN to other points l times. Since n_i are fixed, the covariances of the cell counts can also be obtained as

$$\mathrm{Cov}[S_i, S_j] = \mathrm{Cov}(N_{ii}, N_{jj}) = (n^2-3n-Q+R)p_{iijj} - n^2p_{ii}p_{jj}$$

and

$$\mathrm{Cov}[M_i, M_j] = \mathrm{Cov}(n_i - N_{ii}, n_j - N_{jj}) = \mathrm{Cov}(N_{ii}, N_{jj})$$

where $p_{iijj} = \frac{n_i(n_i-1)n_j(n_j-1)}{n(n-1)(n-2)(n-3)}$. The covariance of cell counts in different columns is

$$\mathrm{Cov}[S_i, M_j] = \begin{cases} \mathrm{Cov}[N_{ii}, n_i - N_{ii}] = -\mathrm{Var}[N_{ii}] & \text{if } i = j, \\ \mathrm{Cov}[N_{ii}, n_j - N_{jj}] = -\mathrm{Cov}[N_{ii}, N_{jj}] & \text{if } i \neq j. \end{cases} \quad (4)$$

See Dixon (1994, 2002a) for the derivation of the above variance and covariance terms.

In a CCT, deviations of S_i or M_i from their expected values under RL or CSR independence can be assessed. Since $S_i = N_{ii}$, for cell $(i,1)$ of the CCT, we have

$$Z_{S_i} = \frac{S_i - \mathrm{E}[S_i]}{\sqrt{\mathrm{Var}[S_i]}} = \frac{N_{ii} - \mathrm{E}[N_{ii}]}{\sqrt{\mathrm{Var}[N_{ii}]}} \quad (5)$$

for $i = 1,2,\ldots,k$. Notice that $Z_{S_i} = Z_{ii}$ where Z_{ii} is the cell-specific tests for cell (i,i) in the NNCT analysis (see, Dixon, 1994 and Ceyhan, 2008a for more details). Notice also that the mixed column entries carry the same information as the self column entries, and they will yield the test statistic with negative sign. That is, $(M_i - \mathrm{E}[M_i])/\sqrt{\mathrm{Var}[M_i]} = -Z_{S_i}$ for each i, hence the test statistics with mixed column entries are omitted. For large n_i, Z_{S_i} approximately has $N(0,1)$ distribution (Dixon, 2002a).

The test statistics for the self cells of the CCT are as follows: For artificial data set 1, we have $Z_{S_1} = Z_{11} = Z_{S_2} = Z_{22} = Z_{S_3} = Z_{33} = 1.4409$ which is in agreement with our observation in the CCT that all classes exhibit mild segregation. For artificial data set 2, $Z_{S_1} = Z_{11} = -1.8033$, $Z_{S_2} = Z_{22} = 1.7388$, and $Z_{S_3} = Z_{33} = -1.5081$ which is in agreement with our observation in the CCT that classes 1 and 3 exhibit lack of segregation at a moderate level while class 2 exhibits mild level of segregation.

One can combine the S_i values (i.e., the self column in the CCT) into a vector $\mathbf{S} = (S_1, S_2, \ldots, S_k) = (N_{11}, N_{22}, \ldots, N_{kk})$. So $\mathrm{E}[\mathbf{S}]$ is the vector of expected values of the entries of $\mathbf{S}$. The variance-covariance matrix of $\mathbf{S}$, denoted $\Sigma_{\mathbf{S}}$, is the $k \times k$ matrix with entry (i,i) being $\mathrm{Var}[S_i] = \mathrm{Var}[N_{ii}]$ and entry (i,j) with $i \neq j$ being $\mathrm{Cov}[S_i, S_j] = \mathrm{Cov}[N_{ii}, N_{jj}]$. With the self column as the vector $\mathbf{S}$, we have the quadratic form

$$\mathscr{X}_C = (\mathbf{S} - \mathrm{E}[\mathbf{S}])^{\intercal} \Sigma_{\mathbf{S}}^{-1} (\mathbf{S} - \mathrm{E}[\mathbf{S}]). \tag{6}$$

where $\Sigma_{\mathbf{S}}^{-1}$ is the inverse of $\Sigma_{\mathbf{S}}$. For large n_i, $\mathscr{X}_C$ approximately has a χ^2_k distribution. Observe that the test statistic $\mathscr{X}_C$ is obtained similar to the overall segregation test as described in Ceyhan (2008a). Briefly, the overall segregation test due to Dixon is

$$\mathscr{X}_D = (\mathbf{N} - \mathrm{E}[\mathbf{N}])^{\intercal} \Sigma_{\mathbf{N}}^{-} (\mathbf{N} - \mathrm{E}[\mathbf{N}]) \tag{7}$$

where $\mathbf{N}$ is the vector of entries of NNCT concatenated row-wise and $\Sigma_{\mathbf{N}}$ is the covariance matrix of $\mathbf{N}$ and A^- is the generalized inverse of a matrix A (Searle, 2006).

For the artificial data set 1, we have $\mathscr{X}_C = 5.0761$ $(p = 0.1664)$ and $\mathscr{X}_D = 7.1274$ $(p = 0.3092)$. Notice that neither test is significant, although the correspondence test yields a lower p-value. This suggests lack of significant deviations from the expected cell counts in either contingency table. On the other hand, for the artificial data set 2, we have $\mathscr{X}_C = 9.4670$ $(p = 0.0237)$ and $\mathscr{X}_D = 9.7879$ $(p = 0.1339)$. Notice that the overall segregation test is not significant at the .05 level, which suggests that the cell counts do not deviate significantly from their expected values. On the other hand, $\mathscr{X}_C$ is significant, which is suggesting significant deviation in the first column of CCT (or the diagonal of NNCT). However, to determine the direction of correspondence, we assess the cell counts in the CCT and conclude that there is an abundance of self pairs for class 2, while there is a lower number of self pairs (or there is an abundance of mixed pairs) for the other classes. Together with the column sums in the CCT, we observe that there is evidence for mixed correspondence compared to self correspondence.

Alternatively, we could also concatenate self and mixed columns of CCT to obtain the vector $\mathbf{S}_{II} = (N_{11}, N_{22}, \ldots, N_{kk}, n_1 - N_{11}, n_2 - N_{22}, \ldots, n_k - N_{kk})$ with the test statistic $\mathscr{X}_{II} = (\mathbf{S}_{II} - \mathrm{E}[\mathbf{S}_{II}])' \Sigma_{II}^{-} (\mathbf{S}_{II} - \mathrm{E}[\mathbf{S}_{II}])$, but this version is highly unstable due to severe rank deficiency (see Ceyhan, 2014). Thus we employ the first form of the test statistic, $\mathscr{X}_C$, which is the χ^2 test for the self column and omit $\mathscr{X}_{II}$ in our further discussion.

When $\mathscr{X}_C$ is significant, it implies the presence of significant deviation of some of the cell counts S_i than expected under H_o in Equation (9) or small deviations of cell counts in positive or negative direction might accumulate in the quadratic form in Equation (6) and cause a significant result for $\mathscr{X}_C$. Furthermore, if some significant deviation exists for some cell(s), this deviation could be toward significant segregation or lack of segregation for a class, or significant association of this class with some other class(es). If additionally, the deviations of cells are all toward positive direction (i.e., segregation) or deviations of some cells toward segregation are strong enough, then the self pairs might be more abundant indicating presence of self correspondence. So with $\mathscr{X}_C$ to infer self or mixed correspondence, one needs to check the direction and magnitude of deviation for each class (after a significant $\mathscr{X}_C$), hence should look at the sign and magnitude of the cell-specific Z tests (i.e., the diagonal cell-specific tests) in Equation (5). Thus this process tests self or mixed correspondence by a two-step approach which may be somewhat a subjective assessment of magnitude of the deviations. For example, in our artificial data set 2, the correspondence test statistic is significant, but by itself, does not indicate it is self or mixed correspondence. To determine the type of correspondence, we either look at the CCT or the sign and magnitude of the tests for the cells in the self column of the CCT. In particular, in this data set, we observe that the correspondence is of mixed type due to large negative values for the Z_{S_1} and Z_{S_3}.

As an alternative approach, we propose a test based on the sum of the self column, S, in the CCT. That is,

$$Z_C = \frac{S - \mathrm{E}[S]}{\sqrt{\mathrm{Var}[S]}}. \tag{8}$$

Here

$$\mathrm{E}[S] = \mathrm{E}\left[\sum_{i=1}^{k} N_{ii}\right] = \sum_{i=1}^{k} \mathrm{E}[N_{ii}] = \sum_{i=1}^{k} \frac{n_i(n_i-1)}{n-1}$$

and

$$\mathrm{Var}[S] = \mathrm{Var}\left[\sum_{i=1}^{k} N_{ii}\right] = \sum_{i=1}^{k} \mathrm{Var}[N_{ii}] + \sum_{i \neq j}^{k} \mathrm{Cov}[N_{ii}, N_{jj}].$$

Observe that $\mathrm{Var}[S]$ is the sum of entries of Σ_{self}, the covariance matrix of $\mathbf{S}$. As n_i values tend to infinity, Z_C converges in law to $N(0,1)$ distribution. Large (positive) values of Z_C indicate that self pairs are more abundant than expected under RL or CSR independence, hence indicate presence of self correspondence, while smaller (negative) values of Z_C indicate presence of mixed correspondence.

For artificial data set 1, $Z_C = 2.2529$ ($p = 0.0123$) which indicates that the self column sum is significantly larger than its expected value. Since each cell count deviates in the same direction, this constitutes evidence for self correspondence in the NN structure. Notice that although segregation is mild (and not significant) for each class, their cumulative effect makes the number of self NN pairs significantly higher than expected yielding a significant self correspondence. As for artificial data set 2, $Z_C = -0.8137$ which implies the self column sum in the CCT is not significantly different from its expected value. However, this is not a contradiction with our finding of significant $\mathscr{X}_C$, as the deviations in the first column are in opposite directions, hence cancel each other out in the summation.

Although the test statistics, $\mathscr{X}_C$, Z_{ii}, and Z_C are all related to correspondence and segregation, they test different null hypotheses. The null hypothesis for correspondence is

$$H_o : \text{self (or mixed) NN pairs are as expected under RL and CSR independence.} \quad (9)$$

Hence, by construction, the cell-specific test Z_{ii} tests the hypothesis

$$H_o : \mathrm{E}[Z_{ii}] = \frac{n_i(n_i-1)}{n-1} \quad (10)$$

and $\mathscr{X}_C$ tests the hypothesis

$$H_o : \mathrm{E}[Z_{ii}] = \frac{n_i(n_i-1)}{n-1} \text{ for all } i = 1,2,\ldots,k \quad (11)$$

and Z_C tests the hypothesis

$$H_o : \mathrm{E}[S] = \sum_{i=1}^{k} \frac{n_i(n_i-1)}{n-1}. \quad (12)$$

The right (resp. left) sided alternative for H_o in Equation (12) will imply self (resp. mixed) correspondence, and the right sided alternative for H_o in Equation (10) will imply segregation of species i from others. On the other hand, the left sided alternative for H_o in Equation (10) will imply lack of segregation of species i from others (in a two-class setting, this is equivalent to association of the species with the other, but in a multi-class setting, this may or may not imply association).

For $k = 2$ classes, $\mathscr{X}_C$ is equivalent to the overall test of segregation of Dixon (1994), $\mathscr{X}_D$, since the CCT and NNCT convey the same information and both tests are effectively based on N_{11} and N_{22} only. In particular, N_{11} and N_{22} constitute the first column of the CCT and the diagonal entries of the NNCT and N_{12} and N_{21} constitute the second column of the CCT and the off-diagonal entries of the NNCT. But for $k > 2$, the information conveyed by the NNCT and CCT are different and the $\mathscr{X}_C$ depends only on

$S_i = N_{ii}$ values in CCT, while the overall segregation test depends on all N_{ij} values in NNCT.

Remark 3.2. Relation of Null Hypotheses with CSR Independence and RL. The above null hypotheses in Equation (10)-(12) in terms of the expected values can result from a more general setting. In particular, these null cases follow provided that there is randomness in the NN structure in such a way that the probability of a NN of a point being from a class is proportional to the relative frequency of that class. This assumption holds, e.g., under CSR independence or RL of the points from each class. Both CSR independence and RL patterns imply that there is no correspondence in the NN structure. In fact, it is conceivable that other independence patterns (in which all classes are independently generated from the same process or distribution) can yield the same null hypothesis, but we restrict our attention to RL and CSR independence as they are considered to be the benchmark patterns in spatial data analysis. □

Remark 3.3. Status of Q and R under RL and CSR independence. Note the status of the quantities Q and R under CSR independence and RL models. Under RL, Q and R are fixed, while, under CSR independence, they are random. Hence the tests in Equations (5)-(8) are conditional on the observed values of Q and R under CSR independence while no such conditioning is required under RL. The variance and covariance terms in Section 3.2 and all the corresponding tests also depend on Q and R. Hence these expressions are appropriate for the RL pattern, but for the CSR independence pattern, they are variances and covariances conditioned on Q and R. The unconditional variances and covariances can be obtained by replacing Q and R with their expectations. Under HPP in the infinite plane, Cox (1981) computed $\mathrm{E}[R/n] \to .6215$ and Cuzick and Edwards (1990) computed $\mathrm{E}[Q/n] \to .633$ as $n \to \infty$. However, these results are assuming an infinite plane, and our CSR independence case requires a bounded support (e.g., the unit square) and fixed number of points which renders their computation for exact and asymptotic settings an arduous task (due to, e.g., the edge effects). Alternatively, the expected values of Q and R can be empirically approximated and used in the expressions. For example, for the binomial process on the unit square, $\mathrm{E}[Q/n]$ tends approximately to .6324 and $\mathrm{E}[R/n]$ tends approximately to 0.6219 (estimated empirically based on 1000000 Monte Carlo simulations for increasing values of n). Notice that these estimates are pretty close to the results under HPP. Hence one could also replace Q and R with $0.63\,n$ and $0.62\,n$, respectively and obtain the so-called *QR-adjusted* tests but we use the observed values of Q and R in computing our test statistics even when assessing their behavior under CSR independence. As shown in Ceyhan (2008b), QR-adjustment does not improve on the unadjusted NNCT-tests. □

Remark 3.4. Recommended Strategy for $k > 2$ Classes. In the multi-class case with $k > 2$, we recommend the following strategy for the practical implementation of the corresponding tests: Perform $\mathscr{X}_C$ and Z_C to check presence of self or mixed correspondence

or any deviation in the self column and then perform the cell-specific tests to determine which species (if any) exhibit segregation or lack of it. □

3.3. Consistency of Tests

A reasonable test should have more power as the sample size increases, so, we prove the consistency of the tests in question under appropriate hypotheses. Let $\chi^2_{\nu,\alpha}$ be the $100\alpha^{\text{th}}$ percentile of χ^2 distribution with ν degrees of freedom.

Theorem 3.5. *Let the CCT be constructed from completely mapped spatial data under RL. Then*

(i) the one-sided (hence the two-sided) cell-specific tests using Z_{ii} given in Equation (5) *rejecting H_o in Equation* (10) *are consistent,*

(ii) the test rejecting H_o in Equation (11) *for $\mathscr{X}_C > \chi^2_{k,1-\alpha}$ with $\mathscr{X}_C$ as in Equation* (6) *is consistent,*

(iii) the one-sided (hence the two-sided) tests using Z_C given in Equation (8) *rejecting H_o in Equation* (12) *are consistent.*

Proof. (i) In the k class case, let $T_{n,i} = \frac{S_i/n - \mathrm{E}[S_i/n]}{\sqrt{\mathrm{Var}[S_i/n]}} = \frac{N_{ii}/n - \mathrm{E}[N_{ii}/n]}{\sqrt{\mathrm{Var}[N_{ii}/n]}}$, then $T_{n,i} = Z_{ii}$ for $i = 1,2,\ldots,k$. Consistency of Z_{ij} was proved in Ceyhan (2010) for all i, j which includes the special case of $i = j$, but we still present it here for the sake of completeness. Under RL, $\mathrm{E}[T_{n,i}] = \mathrm{E}[Z_{ii}] = 0$ and $Z_{ii} = (N_{ii} - \mathrm{E}[N_{ii}])/\sqrt{\mathrm{Var}[N_{ii}]}$ are approximately distributed as $N(0,1)$ for large n_i for $i = 1,2,\ldots,k$. Under the right sided (resp. left sided) alternative H_a, for any $i \in \{1,2,\ldots,k\}$, we have $\mathrm{E}[Z_{ii}|H_a] = \varepsilon_i > 0$ (resp. $\mathrm{E}[Z_{ii}|H_a] = \varepsilon_i < 0$) where ε_i is a parameterization of the alternative for class i for $i = 1,2,\ldots,k$. Let $R(\varepsilon_i)$ and $Q(\varepsilon_i)$ be the numbers of reflexive pairs and shared NNs, respectively, $p_{ii}(\varepsilon_i)$, $p_{iii}(\varepsilon_i)$, and $p_{iiii}(\varepsilon_i)$ be the counterparts of p_{ii}, p_{iii}, and p_{iiii} in Equation (3). Then under H_a, we have $\mathrm{Var}[N_{ii}/n] = (1/n + R(\varepsilon_i)/n^2)p_{ii}(\varepsilon_i) + (2/n - 2R(\varepsilon_i)/n^2 + Q(\varepsilon_i)/n^2)p_{iii}(\varepsilon_i) + (1 - 3/n - Q(\varepsilon_i)/n^2 + R(\varepsilon_i)/n^2)p_{iiii}(\varepsilon_i) - (p_{ii}(\varepsilon_i))^2$. So, under H_a, it follows that $\mathrm{Var}[N_{ii}/n] \to 0$ as $n_i \to \infty$. Hence the test using Z_{ii} is consistent for the right-sided (resp. left sided) alternative. Consistency for the two-sided alternative follows similarly.

(ii) Let $\vec{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_k)$, then we have $H_a : \vec{\varepsilon} \neq \mathbf{0}$, with $\mathbf{0}$ being the vector of k zeros. Also let $\lambda(\vec{\varepsilon})$ be the non-centrality parameter of χ^2_k distribution for $\mathscr{X}_C$ under H_a. The α-level test based on $\mathscr{X}_C$ is consistent, since $\mathscr{X}_C$ is a quadratic form based on Z_{ii} values, i.e., $\mathscr{X}_C \sim \chi^2_k(\lambda(\vec{\varepsilon}))$ for some $\lambda(\vec{\varepsilon}) > 0$. Furthermore, for large n, the null and alternative hypotheses are equivalent to $H_o : \lambda = 0$ versus $H_a : \lambda = \lambda(\vec{\varepsilon}) > 0$. Then by standard arguments for the consistency of χ^2 tests, consistency follows.

(iii) Let $T_{n,sc} = \frac{S/n - \mathrm{E}[S/n]}{\sqrt{\mathrm{Var}[S/n]}} = \frac{\sum_{i=1}^k N_{ii}/n - \mathrm{E}\left[\sum_{i=1}^k N_{ii}/n\right]}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^k N_{ii}/n\right]}}$, then $T_{n,sc} = Z_C$. Under RL, $\mathrm{E}[T_{n,sc}] = \mathrm{E}[Z_C] = 0$ and $Z_C = (S - \mathrm{E}[S])/\sqrt{\mathrm{Var}[S]}$ is approximately distributed as $N(0,1)$

for large n. Under right-sided (resp. left sided) alternative H_a, we have $\mathrm{E}[S|H_a] = \varepsilon > 0$ (resp. $\mathrm{E}[S|H_a] = \varepsilon < 0$) where ε is a parameterization of the alternative with $\varepsilon > 0$ (resp. $\varepsilon < 0$) characterizing self (resp. mixed) correspondence. Let $R(\varepsilon)$ and $Q(\varepsilon)$ be the numbers of reflexive pairs and shared NNs, respectively, $p_{ii}(\varepsilon)$, $p_{iii}(\varepsilon)$, and $p_{iiii}(\varepsilon)$ be the counterparts of p_{ii}, p_{iii}, and p_{iiii} in Equation (3). Then under H_a, we have $\mathrm{Var}[N_{ii}/n] = (1/n + R(\varepsilon)/n^2)p_{ii}(\varepsilon) + (2/n - 2R(\varepsilon)/n^2 + Q(\varepsilon)/n^2)p_{iii}(\varepsilon) + (1 - 3/n - Q(\varepsilon)/n^2 + R(\varepsilon)/n^2)p_{iiii}(\varepsilon) - (p_{ii}(\varepsilon))^2$ and $\mathrm{Cov}[N_{ii}/n, N_{jj}/n] = (1 - 3/n - Q(\varepsilon)/n^2 + R(\varepsilon)/n^2)p_{iijj} - p_{ii}p_{jj}$. So, under H_a, it follows that $\mathrm{Var}[N_{ii}/n] \to 0$ and $\mathrm{Cov}[N_{ii}/n, N_{jj}/n] \to 0$ as $n_i \to \infty$. Hence $\mathrm{Var}[S] \to 0$ as $n_i \to \infty$. Thus the test using Z_C is consistent for the right-sided (resp. left sided) alternative. Consistency for the two-sided alternative is similar. ■

4. Empirical Size and Power Analysis

In this section we investigate the finite sample performance of the tests under RL or CSR independence and under various alternatives via Monte Carlo simulations.

4.1. Empirical Size Analysis

To determine empirical size performance of the tests, we use CSR independence and RL as our null hypotheses. Under these patterns, correspondence would occur at expected levels. That is, under these patterns we have $\mathrm{E}[S_i] = n_i(n_i - 1)/(n - 1)$ for all $i = 1, 2, \ldots, k$ as in Equation (11) and $\mathrm{E}[S] = \sum_{i=1}^{k} n_i(n_i - 1)/(n - 1)$ as in Equation (12).

We estimate the empirical sizes (i.e., significance levels) based on the asymptotic critical values. For example, let T be a test with a χ^2_{df} distribution asymptotically, and let T_i be the value of test statistic for the sample generated at i^{th} Monte Carlo replication for $i = 1, 2, \ldots, N_{mc}$. Then the empirical size of T at level $\alpha = 0.05$, denoted $\widehat{\alpha}_T$ is computed as $\widehat{\alpha}_T = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}\left(T_i \geq \chi^2_{df,0.95}\right)$. Furthermore, let Z be a test with a $N(0,1)$ asymptotic distribution, and let Z_i be the value of test statistic for i^{th} sample generated. Then the empirical size of Z for the left-sided (resp. right-sided) alternative at level $\alpha = 0.05$, denoted $\widehat{\alpha}_Z$ is computed as $\widehat{\alpha}_Z = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}(Z_i \leq z_{0.05} = -1.645)$ (resp. $\widehat{\alpha}_Z = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}(Z_i \geq z_{0.95} = 1.645)$). The empirical size for the two-sided alternative is computed as $\widehat{\alpha}_Z = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}(|Z_i| \geq z_{0.975} = 1.96)$.

4.1.1. Empirical Size Analysis under CSR Independence

We consider the two-class and three-class cases. For the three-class case, we have classes X, Y, and Z (or classes 1, 2, and 3) of sizes n_1, n_2, and n_3 respectively. Under H_o, at each of $N_{mc} = 10000$ replications, we generate n_1 X points $\mathscr{X}_{n_1} = \{X_1, \ldots, X_{n_1}\}$,

n_2 Y points $\mathscr{Y}_{n_2} = \{Y_1, \ldots, Y_{n_2}\}$, and n_3 Z points $\mathscr{Z}_{n_3} = \{Z_1, \ldots, Z_{n_3}\}$ independently of each other and iid from $\mathscr{U}((0,1) \times (0,1))$ and combine X, Y and Z points as $\mathscr{W}_n = \mathscr{X}_{n_1} \cup \mathscr{Y}_{n_2} \cup \mathscr{Z}_{n_3} = \{W_1, W_2, \ldots, W_n\}$. For the two-class case, we only generate points from classes X and Y and combine them as $\mathscr{W}_n = \mathscr{X}_{n_1} \cup \mathscr{Y}_{n_2} = \{W_1, W_2, \ldots, W_n\}$. We consider four cases for CSR independence:

$$\text{CSR Case 1 (with 2 classes)} : n_1 = n_2 = n = 10, 20, 30, 40, 50$$
$$\text{CSR Case 2 (with 2 classes)} : n_1 = 20 \text{ and } n_2 = 20, 30, \ldots, 60.$$
$$\text{CSR Case 3 (with 3 classes)} : n_1 = n_2 = n_3 = n = 10, 20, 30, 40, 50$$
$$\text{CSR Case 4 (with 3 classes)} : n_1 = 20,\ n_2 = 40, \text{ and } n_3 = 40, 40, \ldots, 80.$$

In CSR cases 1 and 3, the sample sizes are equal and increasing, to determine the influence of the increasing balanced sample sizes on the empirical levels of the tests. On the other hand, CSR cases 2 and 4 are designed to determine the influence of differences in the sample sizes (i.e., differences in relative abundances of classes) on the empirical levels of the tests.

The empirical significance levels for the tests under CSR independence are presented in Table 6, where $\widehat{\alpha}_{Z_{11}}$, $\widehat{\alpha}_{Z_{22}}$, and $\widehat{\alpha}_{Z_{33}}$ are for the cell-specific tests for cells $(1,1)$, $(2,2)$, and $(3,3)$ (for segregation); (see, e.g., Dixon, 1994 and Ceyhan, 2008a for details on the cell-specific tests); $\widehat{\alpha}_{\mathscr{X}_C}$ is for the χ^2 test $\mathscr{X}_C$, testing the self column in CCT; $\widehat{\alpha}_{Z_C}$ is for Z_C, testing the sum of the self column; $\widehat{\alpha}_{\mathscr{X}_D}$ is for Dixon's overall segregation test; and $\widehat{\alpha}_{\mathscr{X}_C,Z_C}$ is the proportion of agreement in rejecting the null hypothesis for $\mathscr{X}_C$ and Z_C; $\widehat{\alpha}_{\mathscr{X}_D,\mathscr{X}_C}$ is the proportion of agreement for $\mathscr{X}_D$ and $\mathscr{X}_C$; and $\widehat{\alpha}_{\mathscr{X}_D,Z_C}$ is the proportion of agreement for $\mathscr{X}_D$ and Z_C. For $N_{mc} = 10000$ replications, an empirical size estimate is deemed conservative, if smaller than .0464 while it is deemed liberal, if larger than .0536 at .05 level (based on binomial critical values with $n = 10000$ trials and probability of success 0.05).

In the two-class cases (i.e., CSR cases 1 and 2), we do not present Dixon's overall test of segregation as it is identical to $\mathscr{X}_C$ for two classes. Under CSR case 1, $\mathscr{X}_C$ and Z_C are slightly conservative for smaller sample sizes. Under CSR case 2, $\mathscr{X}_C$ and Z_{11} are conservative (with the latter being more so) when sample sizes are unbalanced (i.e., the relative abundance ratio, n_2/n_1, gets larger than two). Note also that Z_C seems to be robust to differences in relative abundance of the classes. The proportion of agreement in rejecting the null hypothesis by $\mathscr{X}_C$ and Z_C is significantly smaller than .05, which implies these tests have significantly different rejection/acceptance regions (i.e., they are testing substantially different hypotheses).

Under the three-class cases of CSR cases 3 and 4, we also present Dixon's overall test of segregation as it is different from $\mathscr{X}_C$ for more than two classes. Under CSR case 3, all tests are slightly conservative for smaller sample sizes and cell-specific tests are slightly liberal for larger sample sizes. Under CSR case 4, Z_{11} is conservative for all sample size combinations (since it has the smallest sample size in this case where

Table 6: *The empirical significance levels of the tests under CSR independence cases 1-4 with $N_{mc} = 10000$ at $\alpha = .05$. $\widehat{\alpha}_{Z_{11}}$, $\widehat{\alpha}_{Z_{22}}$, and $\widehat{\alpha}_{Z_{33}}$ are the empirical significance levels for the cell-specific tests for cells $(1,1)$, $(2,2)$, and $(3,3)$ (for segregation); $\widehat{\alpha}_{\mathscr{X}_D}$ for Dixon's overall segregation test, $\mathscr{X}_D$; $\widehat{\alpha}_{\mathscr{X}_C}$ for the χ^2 test $\mathscr{X}_C$; $\widehat{\alpha}_{Z_C}$ for Z_C; and $\widehat{\alpha}_{\mathscr{X}_C,Z_C}$ is the proportion of agreement in rejecting the null hypothesis for $\mathscr{X}_C$ and Z_C; $\widehat{\alpha}_{\mathscr{X}_D,\mathscr{X}_C}$ is the proportion of agreement for $\mathscr{X}_D$ and $\mathscr{X}_C$; and $\widehat{\alpha}_{\mathscr{X}_D,Z_C}$ is the proportion of agreement for $\mathscr{X}_D$ and Z_C. Size estimates larger than .0536 (resp. smaller than .0464) are liberal (resp. conservative) and are superscripted with $^{\ell}$ (resp. ^{c}).*

CSR case 1					
n	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{\mathscr{X}_C,Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$
10	$.0432^{c}$	$.0439^{c}$	.0216	$.0454^{c}$	.0465
20	$.0457^{c}$	$.0443^{c}$	.0207	.0517	.0522
30	.0485	$.0462^{c}$	.0237	.0573	.0493
40	.0501	$.0545^{\ell}$	.0254	.0507	.0525
50	.0472	.0468	.0215	$.0454^{c}$	.0472

CSR case 2					
n_2	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{\mathscr{X}_C,Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$
20	$.0437^{c}$	$.0448^{c}$	.0197	.0482	.0517
30	.0480	.0493	.0253	.0521	.0479
40	.0489	.0521	.0237	$.0313^{c}$	$.0455^{c}$
50	$.0427^{c}$	.0526	.0219	$.0295^{c}$	.0478
60	$.0452^{c}$	.0465	.0233	$.0395^{c}$	.0495

CSR case 3									
n	$\widehat{\alpha}_{\mathscr{X}_D}$	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{\mathscr{X}_D,\mathscr{X}_C}$	$\widehat{\alpha}_{\mathscr{X}_D,Z_C}$	$\widehat{\alpha}_{\mathscr{X}_C,Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$	$\widehat{\alpha}_{Z_{33}}$
10	$.0421^{c}$	$.0425^{c}$	.0491	.0179	.0084	.0312	$.0277^{c}$	$.0283^{c}$	$.0250^{c}$
20	$.0408^{c}$	$.0438^{c}$	.0481	.0180	.0094	.0293	$.0332^{c}$	$.0283^{c}$	$.0318^{c}$
30	.0465	.0473	.0496	.0204	.0110	.0320	.0530	.0526	$.0549^{\ell}$
40	$.0455^{c}$	.0495	$.0461^{c}$	.0205	.0092	$.0320^{c}$	.0509	$.0558^{\ell}$	$.0595^{\ell}$
50	.0474	.0497	.0504	.0229	.0120	.0329	$.0605^{\ell}$	$.0588^{\ell}$	$.0564^{\ell}$

CSR case 4									
n_3	$\widehat{\alpha}_{\mathscr{X}_D}$	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{\mathscr{X}_D,\mathscr{X}_C}$	$\widehat{\alpha}_{\mathscr{X}_D,Z_C}$	$\widehat{\alpha}_{\mathscr{X}_C,Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$	$\widehat{\alpha}_{Z_{33}}$
40	.0490	.0509	.0492	.0233	.0126	.0342	$.0418^{c}$	$.0551^{\ell}$	.0510
50	$.0412^{c}$	$.0443^{c}$	$.0450^{c}$	.0187	.0100	.0297	$.0344^{c}$	$.0460^{c}$	.0489
60	.0488	.0466	.0528	.0212	.0123	.0354	$.0238^{c}$	$.0543^{\ell}$	.0492
70	.0528	.0496	.0498	.0261	.0156	.0344	$.0458^{c}$	.0520	.0517
80	.0509	.0492	.0518	.0228	.0116	.0302	$.0333^{c}$	$.0431^{c}$	.0522

there is substantial class imbalance) whereas Z_{33} has the best size performance as it corresponds to the class with the largest samples. The proportions of agreement by the tests in rejecting the null hypothesis are all significantly smaller than .05, which implies these tests have significantly different rejection/acceptance regions (with $\mathscr{X}_C$ and Z_C having the largest overlap (i.e., these statistics are testing more similar hypotheses) and $\mathscr{X}_D$ and Z_C having the smallest overlap in rejection regions (i.e., these statistics are testing more different hypotheses compared to other pairs).

For unbalanced or small sample sizes, the tests are usually conservative (especially for the cell-specific tests for the smaller samples), so we recommend the use of the Monte Carlo randomized versions or the use of Monte Carlo critical values for the cell-specific test for the smaller class. A Monte Carlo critical value is determined as the appropriately ranked value of the test statistic in a certain number of generated data sets

from the distribution under the null hypothesis. The class sizes are said to be *balanced*, if the relative abundances of the classes are close to one, and they are called *unbalanced*, if the relative abundances deviate substantially from one.

4.1.2. Empirical Size Analysis under RL

Under the RL pattern, the class labels or marks are assigned randomly to points whose locations are given. Recall that $\mathscr{W}_n = \{w_1, w_2, \ldots, w_n\}$ is the given set of locations for n points from the background pattern. For two classes, at each background realization, n_1 of the points are labeled as class 1 or X and the remaining $n_2 = n - n_1$ points are labeled as class 2 or Y. Similarly, for three classes, at each background realization, n_1 of the points are labeled as class X, n_2 of the points are labeled as class Y, and the remaining $n_3 = n - (n_1 + n_2)$ points are labeled as class Z.

Types of the Background Patterns (Two Classes)

RL Case 1: The background points are a realization of $Z_i \overset{iid}{\sim} \mathscr{U}((0,1) \times (0,1))$ for $i = 1, 2, \ldots, n$. That is, the background points, $\mathscr{W}_n$, are generated iid uniform in the unit square $(0,1) \times (0,1)$. We consider $n_1 = n_2 = 10, 20, \ldots, 50$.

RL Case 2: The background points, $\mathscr{W}_n$, are generated as in case 1 above with $n_1 = 20$ and $n_2 = 20, 30, \ldots, 60$.

RL Case 3: The background points, $\mathscr{W}_n$, are generated from a Matérn cluster process, MatClust(κ, r, μ) (Baddeley and Turner, 2005). In this process, first "parent" points are generated from a Poisson process with intensity κ. Then each parent point is replaced by $N \sim$ Poisson(μ) new points which are generated iid inside the circle of radius r centered at the parent point. Each background realization is a realization of $\mathscr{W}_n$ and is generated from MatClust(κ, r, μ). Let n be the number of points in a particular realization. Then $n_1 = \lfloor n/2 \rfloor$ of these points are labeled as class 1 where $\lfloor x \rfloor$ stands for the floor of x, and $n_2 = n - n_1$ as class 2. In our simulations, we use $\kappa = 2, 4, \ldots, 10$, $\mu = \lfloor 100/\kappa \rfloor$, and $r = 0.1$. That is, we take $(\kappa, \mu) \in \{(2,50), (4,25) \ldots, (10,10)\}$ so as to have about 100 background points on the average with about half of them being from class 1 and the other half being from class 2.

To reduce the influence of a particular background realization on the size performance of the tests, we generate 100 different realizations of each background pattern. For each case, the RL scheme described is repeated 1000 times for each (n_1, n_2) combination at each of 100 different background realizations. So we have $N_{mc} = 100000$. In RL cases 1 and 2, the points are from HPP in the unit square with fixed n_1 and n_2 (i.e., from binomial process), where RL case 1 is for assessing the effect of equal but increasing sample sizes on the tests, while RL case 2 is for assessing the effect of increasing differences in sample sizes of the classes (with one class size being fixed, while the other is

increasing). On the other hand, in the background realizations of RL case 3, centers and numbers of clusters are random. On the average, with increasing κ, the cluster sizes tend to decrease and the number of clusters tend to increase (so as to have fixed class sizes on the average). Hence in RL case 3, we investigate the influence of increasing number of clusters with randomly determined centers on the size performance of the tests.

The empirical size estimates of the tests for two classes under RL cases 1-3 are presented in Table 7. For $N_{mc} = 100000$ replications, if all the Monte Carlo replications were independent, an empirical size estimate would have been deemed conservative, if smaller than .04887 while it would have been deemed liberal, if larger than .05113 at .05 level (based on binomial critical values with $n = 100000$ trials and probability of success 0.05). This approach is like providing critical values for a two-sided hypothesis test. Equivalently, one might construct a confidence interval (say 95 %) for the proportion of rejections (i.e., empirical size estimate) and check whether it contains the nominal level of .05 or lies completely at one side of .05. However, under our RL scheme, the Monte Carlo replications are not independent as 100 replications are performed at each of 100 background realizations, hence within sample independence is violated rending both the critical value and the confidence interval approaches are not appropriate. But we can account for dependence due to the use of same background realization for 100 of the realizations, at each of which 1000 Monte Carlo replications are performed, by using a linear mixed effects model. In particular, in the "lme4" package in R, we can employ "lmer" command with properly declaring the error structure for dependence in the background realization. For example, let "bg" stand for the background factor (i.e., takes the same value for each Monte Carlo replication at the same background realization). Then we can apply a mixed modeling with "lmer" command by declaring the error structure as "$(1|\text{bg})$" and construct a 95 % confidence interval for the size estimate value. We mark the empirical sizes not significantly different from .05 with an asterisk.

Under RL case 1, tests are either slightly conservative or liberal (with more conservative for smaller samples), and under RL case 2, cell-specific tests for the smaller sample is moderately conservative, and the other tests are slightly conservative or liberal. The tests have sizes about the nominal level under RL case 3, since in this case, the class sizes are about 50, which seems large enough for the normal approximation to take effect. Moreover, the size performance of the tests does not depend on the number and size of the clusters in the background pattern and the more important factor is the sample sizes.

Types of the Background Patterns (Three Classes)

RL Case (i): Same as in RL Case 1 of the two class setting with $n_1 = n_2 = n_3 = 10, 20, \ldots, 50$.

RL Case (ii): Same as in RL Case 2 of the two class setting with $n_1 = 20$, $n_2 = 40$ and $n_3 = 40, 50, \ldots, 80$.

RL Case (iii): Same as in RL Case 3 of the two class setting with $n_1 = n_2 = \lfloor n/3 \rfloor$ points are labeled as classes 1 and 2 and $n_3 = n - (n_1 + n_2)$ as class 3. In our simulations, we use $\kappa = 2,4,\ldots,10$, $\mu = \lfloor 150/\kappa \rfloor$, and $r = 0.1$. That is, we take $(\kappa,\mu) \in \{(2,75),(4,37)\ldots,(10,15)\}$ so as to have about 150 background points on the average with about a third of them being from each of classes 1-3.

Table 7: *The empirical significance levels of the tests for two classes under RL cases 1-3 with* $N_{mc} =$ 100000 *(1000 replications for each of 100 background realizations) at* $\alpha = .05$*. The empirical size labeling for the tests is as in Table 6. Size estimates not significantly different from .05 are marked with an asterisk.*

RL case 1				
n	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$
10	.04281	.04276	.04513	.04625
20	.04511	.04612	.05349	.05209
30	.04862*	.04616	.05220	.05258
40	.04782	.05398	.05232	.05217
50	.04942*	.04932*	.04740	.04642

RL case 2				
n_2	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$
20	.04602	.04670	.05479	.05414
30	.04735	.04783	.05050*	.04886*
40	.04551	.05357	.03375	.04358
50	.04611	.05649	.03456	.04893*
60	.04395	.04670	.04042	.04749

RL case 3				
κ	$\widehat{\alpha}_{\mathscr{X}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$
2	.04700	.04957*	.04734	.04577
4	.04804	.04959*	.04901*	.04860*
6	.04905*	.05023*	.05103*	.04926*
8	.04859	.04983*	.05096*	.04914*
10	.04869*	.05011*	.05042*	.05097*

The empirical size estimates of the tests for three classes under RL cases (i)-(iii) are presented in Table 8. Under all cases $\mathscr{X}_D$ and $\mathscr{X}_C$ are slightly conservative (with the former being more conservative), and Z_C is closest to the nominal level. Under RL case (i) cell-specific tests are conservative for smaller samples, under RL case (ii), cell-specific tests for the smaller samples are conservative, while larger samples are close to the nominal level. Under RL case (iii) all Z tests are at about the desired level.

Based on the empirical size performance of the tests under CSR independence and RL, we observe that the new tests $\mathscr{X}_C$ and Z_C are more appropriate for both balanced or unbalanced sample sizes (with the latter being more robust to the imbalance in class sizes).

4.2. Empirical Power Analysis

To compare the empirical power performance of the tests, we consider various alternative cases with the two and three classes for deviations from the null case in the NN structure. The empirical power estimates are computed at $\alpha = .05$ as in the size estimation in Section 4.1.

Table 8: *The empirical significance levels of the tests with three classes under RL cases (i)-(iii) with* $N_{mc} = 10000$ *at* $\alpha = .05$. *The notation is as in Table 6. Size estimates not significantly different from .05 are marked with an asterisk.*

RL Case (i)						
n	$\widehat{\alpha}_{\mathscr{Z}_D}$	$\widehat{\alpha}_{\mathscr{Z}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$	$\widehat{\alpha}_{Z_{33}}$
10	.03879	.03957	.04964*	.02536	.02585	.02528
20	.04479	.04571	.05043*	.03293	.03364	.03175
30	.04558	.04756	.05151	.05292	.05365	.05370
40	.04628	.04773	.04789	.05328	.05231	.05280
50	.04797	.04877*	.05009*	.05855	.05804	.05823

RL Case (ii)						
n_3	$\widehat{\alpha}_{\mathscr{Z}_D}$	$\widehat{\alpha}_{\mathscr{Z}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$	$\widehat{\alpha}_{Z_{33}}$
40	.04701	.04728	.04674	.04048	.05155*	.04995*
50	.04674	.04736	.05120*	.03374	.04592	.05375
60	.04696	.04578	.05006*	.02355	.05240	.05036*
70	.04870*	.04881*	.04689	.04483	.05102*	.04662
80	.04798	.04970*	.05085*	.03427	.04768	.04747

RL Case (iii)						
κ	$\widehat{\alpha}_{\mathscr{Z}_D}$	$\widehat{\alpha}_{\mathscr{Z}_C}$	$\widehat{\alpha}_{Z_C}$	$\widehat{\alpha}_{Z_{11}}$	$\widehat{\alpha}_{Z_{22}}$	$\widehat{\alpha}_{Z_{33}}$
2	.04692	.04728	.04940*	.05081*	.05069*	.04764*
4	.04650	.04836	.04878*	.04752	.04813*	.04886*
6	.04860	.04900*	.04927*	.04825*	.04878*	.04959*
8	.04743	.04836	.04736	.04994*	.04923*	.04833*
10	.04693	.04791	.04868*	.04918*	.04881*	.05032*

4.2.1. Empirical Power Analysis for Two Classes

For the two classes, we consider five alternative cases.

Case I: For this class of alternatives, we generate $X_i \overset{iid}{\sim} \mathscr{U}((0,1)\times(0,1))$ for $i = 1,\ldots,n_1$ and $Y_j \overset{iid}{\sim} \text{BVN}(1/2,1/2,\sigma_1,\sigma_2,\rho)$ for $j = 1,\ldots,n_2$, where $\text{BVN}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ is the bivariate normal distribution with mean (μ_1,μ_2) and covariance $\begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}$. In our simulations, we set $\sigma_1 = \sigma_2 = \sigma$ and $\rho = 0$. We consider the following three alternatives:

$$H_I^1 : \sigma = 1/5, \quad H_I^2 : \sigma = 2/15, \text{ and } H_I^3 : \sigma = 1/10. \tag{13}$$

The classes 1 and 2 (i.e., X and Y) have different distributions with different local intensities. In particular, X points are a realization of uniform distribution in the unit square, while Y points are clustered around the center of the unit square $(1/2,1/2)$ where the

level of clustering increases as σ decreases. This suggests a high level of niche specificity for Y points around the center of the unit square compared to X points, which in turn implies segregation of Y points from X points. Furthermore, self NN pairs would be more likely to occur compared to mixed NN pairs, hence self correspondence is expected to be observed.

The empirical power estimates under the alternatives, $H_I^1 - H_I^3$, with $n_1 = n_2 = 40$ are presented in Table 9, where $\widehat{\beta}_{\mathscr{X}_C}$ is power estimate for the χ^2 test for the self column, $\mathscr{X}_C$; $\widehat{\beta}_{Z_{11}}$ and $\widehat{\beta}_{Z_{22}}$ are for the cell-specific tests for cells $(1,1)$ and $(2,2)$ (for segregation), and $\widehat{\beta}_{Z_C}$ is for the Z test for the sum of self column, Z_C. Under the case I alternatives, the power estimates increase as σ decreases. In particular, the self column test, $\mathscr{X}_C$, and the right-sided cell-specific tests for cells $(1,1)$ and $(2,2)$ have high power estimates, which indicates segregation of Y points from X points and vice versa. Since segregation occurs for both classes, and $\mathscr{X}_C$ has high power implies self correspondence. Also, the right-sided Z test for the sum of the self column has high power, confirming self correspondence in this case. Notice that the Z_C has the highest power estimates.

Case II: For this type of alternatives, first, we generate $X_i \overset{iid}{\sim} \mathscr{U}((0,1)\times(0,1))$ for $i = 1,2,\ldots,n_1$ and for each $j = 1,2,\ldots,n_2$, we generate Y_j around a randomly picked X_i with probability p in such a way that $Y_j = X_i + R_j(\cos T_j, \sin T_j)^t$ where v^t stands for transpose of the vector v, $R_j \sim \mathscr{U}(0, \min_{i\neq j} d(X_i, X_j))$ and $T_j \sim \mathscr{U}(0, 2\pi)$ or generate Y_j uniformly in the unit square with probability $1-p$. In the pattern generated, Y_j are more associated with X_i. The three values of p constitute the following alternatives:

$$H_{II}^1 : p = .25, \quad H_{II}^2 : p = .50, \text{ and } H_{II}^3 : p = .75. \tag{14}$$

Table 9: *The power estimates under the case I-III, and V alternatives in Equations* (13)-(15)*, and* (17) *with* $N_{mc} = 10000$*,* $n_1 = n_2 = 40$ *at* $\alpha = .05$*.* $\widehat{\beta}_{Z_{11}}$ *and* $\widehat{\beta}_{Z_{22}}$ *are is power estimates for the cell-specific tests for cells* $(1,1)$ *and* $(2,2)$ *(for segregation),* $\widehat{\beta}_{\mathscr{X}_C}$ *is for* $\mathscr{X}_C$*, testing deviations in the self column, and* $\widehat{\beta}_{Z_C}$ *is for* Z_C*, testing the sum of self column. The "*$>$*" (resp. "*$<$*") sign in the superscript implies the power is estimated for the right-sided (resp. left-sided) alternative.*

Power estimates under case I alternatives				
	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{>}_{Z_C}$	$\widehat{\beta}^{>}_{Z_{11}}$	$\widehat{\beta}^{>}_{Z_{22}}$
H_I^1	.2226	.4167	.2648	.3320
H_I^2	.8523	.9599	.8403	.9164
H_I^3	.9929	.9994	.9887	.9972

Power estimates under case II alternatives				
	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{<}_{Z_C}$	$\widehat{\beta}^{<}_{Z_{11}}$	$\widehat{\beta}^{<}_{Z_{22}}$
H_{II}^1	.1469	.3998	.2658	.2330
H_{II}^2	.4051	.7788	.5625	.4054
H_{II}^3	.5393	.9003	.7373	.3366

case III alternatives				
	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{>}_{Z_C}$	$\widehat{\beta}^{>}_{Z_{11}}$	$\widehat{\beta}^{>}_{Z_{22}}$
H_{III}^1	.4196	.6812	.5141	.5134
H_{III}^2	.9247	.9876	.9437	.9439
H_{III}^3	.9999	1.000	.9999	.9997

case V alternatives				
	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{<}_{Z_C}$	$\widehat{\beta}^{<}_{Z_{11}}$	$\widehat{\beta}^{<}_{Z_{22}}$
H_V^1	.1795	.3499	.2160	.3867
H_V^2	.4384	.7081	.5562	.6280
H_V^3	.6808	.8937	.7937	.7795

In this case, X points constitute a realization of the uniform distribution in the unit square, while Y points are clustered around the X points, and the level of clustering increases as p increases. The empirical power estimates under the alternatives, $H_{II}^1 - H_{II}^3$, with $n_1 = n_2 = 40$ are presented in Table 9. Notice that $\mathscr{X}_C$ implies significant deviations in the self column, but Z_{11} and Z_{22} have high power estimates for the left-sided alternative, which implies significant association between the classes. Z_{11} having higher power for the left-sided alternative is due to severe lack of segregation of class X points from class Y points (or class Y points being significantly associated with class X points), and Z_{22} has smaller power since Y points are clustered around X points, which also causes slight clustering of Y points. Furthermore, Z_C has high power for the left-sided alternative, which implies mixed NN pairs are more abundant, hence there is significant mixed correspondence in the NN structure.

Case III: For this class of alternatives, we consider $X_i \overset{iid}{\sim} \mathscr{U}((0,1-s)\times(0,1-s))$ for $i = 1,\ldots,n_1$, and $Y_j \overset{iid}{\sim} \mathscr{U}((s,1)\times(s,1))$ for $j = 1,\ldots,n_2$. The three values of s constitute the following alternatives;

$$H_{III}^1 : s = 1/6, \quad H_{III}^2 : s = 1/4, \text{ and } H_{III}^3 : s = 1/3. \tag{15}$$

Notice that these alternatives are the segregation alternatives considered for Monte Carlo simulations in Ceyhan (2010). The empirical power estimates under the segregation alternatives $H_{III}^1 - H_{III}^3$ are presented in Table 9. The tests have high power which increases as s increases. There is significant segregation (at the same level for both classes by construction), and the cell-specific tests are also significant for the right-sided alternatives. Furthermore, $\mathscr{X}_C$ indicates significant deviations in the self column, and Z_C has high power for the right-sided alternative, indicating self correspondence in the NN structure.

Case IV: We also consider alternatives in which, by construction, self-reflexive pairs are more frequent than expected under CSR independence. We generate $X_i \overset{iid}{\sim} S_1$ for $i = 1,\ldots,\lfloor n_1/2 \rfloor$ and $Y_j \overset{iid}{\sim} S_2$ for $j = 1,\ldots,\lfloor n_2/2 \rfloor$. Then for $k = \lfloor n_1/2 \rfloor + 1,\ldots,n_1$, we generate $X_k = X_{k-\lfloor n_1/2 \rfloor} + r\,(\cos T_j, \sin T_j)^t$ and for $l = \lfloor n_2/2 \rfloor + 1,\ldots,n_2$, we generate $Y_l = Y_{l-\lfloor n_1/2 \rfloor} + r\,(\cos T_j, \sin T_j)^t$ where $r \in (0,1)$ and $T_j \sim \mathscr{U}(0, 2\pi)$. Appropriate small choices of r will yield an abundance of self-reflexive pairs. The three values of r we consider constitute the self-reflexivity alternatives at each support pair (S_1, S_2). Then the nine alternative combinations we consider are given by

$$\begin{aligned}
&(i)\ H_{IV}^1 : S_1 = S_2 = (0,1)\times(0,1), \text{ (a) } r = 1/7, \text{ (b) } r = 1/8, \text{ (c) } r = 1/9,\\
&(ii)\ H_{IV}^2 : S_1 = (0,5/6)\times(0,5/6) \text{ and } S_2 = (1/6,1)\times(1/6,1), \text{ (a) } r = 1/7, \text{ (b) } r = 1/8, \text{ (c) } r = 1/9,\\
&(iii)\ H_{IV}^3 : S_1 = (0,3/4)\times(0,3/4) \text{ and } S_2 = (1/4,1)\times(1/4,1) \text{ (a) } r = 1/7, \text{ (b) } r = 1/8, \text{ (c) } r = 1/9.
\end{aligned} \tag{16}$$

Table 10: *The power estimates under the case IV alternatives with $N_{mc} = 10000$, $n_1 = n_2 = 40$ at $\alpha = .05$. The empirical power labeling and superscripting for "<" and ">" are as in Table 9.*

Power estimates under the case IV alternatives					
	r	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{>}_{Z_C}$	$\widehat{\beta}^{>}_{Z_{11}}$	$\widehat{\beta}^{>}_{Z_{22}}$
H^1_{IV}	1/7	.8671	.9572	.8866	.8848
	1/8	.9377	.9834	.9455	.9436
	1/9	.9735	.9949	.9746	.9743
H^2_{IV}	1/7	.9440	.9885	.9522	.9523
	1/8	.9740	.9949	.9769	.9779
	1/9	.9890	.9982	.9895	.9908
H^3_{IV}	1/7	.9930	.9990	.9925	.9932
	1/8	.9973	.9996	.9967	.9973
	1/9	.9991	.9999	.9984	.9989

In this case, under H^2_{IV} and H^3_{IV}, by construction, there is segregation of the classes due to the choices of the supports. The empirical power estimates under the alternatives $H^1_{IV} - H^3_{IV}$ are presented in Table 10. Notice that the tests all have very high power estimates. Furthermore, there is significant segregation (at the same level for both classes at each alternative by construction), and the cell-specific tests are also significant for the right-sided alternatives. Furthermore, $\mathscr{X}_C$ has high power estimates indicating significant deviation in the self column and Z_C has high power for the right-sided alternative, indicating self correspondence in the NN structure. The power estimates for these tests increase from H^1_{IV} to H^3_{IV} and they also increase as r decreases from (a) to (c) at each (S_1, S_2) combination. Hence the power estimates increase as the levels of segregation increases.

Case V: In this case, first, we generate $X_i \overset{iid}{\sim} \mathscr{U}((0,1) \times (0,1))$ and then generate Y_j as $Y_j = X_i + r(\cos T_j, \sin T_j)^t$ where $r \in (0,1)$ and $T_j \sim \mathscr{U}(0, 2\pi)$. In the pattern generated, appropriate choices of r will cause Y_j and X_i more associated; that is, a Y point will be more likely to be the NN of an X point, and vice versa. The three values of r we consider constitute the three association alternatives;

$$H^1_V : r = 1/4, \quad H^2_V : r = 1/7, \text{ and } H^3_V : r = 1/10. \tag{17}$$

These are also the association alternatives considered for Monte Carlo simulations in Ceyhan (2010).

The empirical power estimates under $H_V^1 - H_V^3$ are presented in Table 9. Notice that $\mathscr{X}_C$ has high power estimates indicating significant deviations in the self column, and the cell-specific tests have high power estimates for the left-sided alternatives indicating presence of significant association of the classes. Furthermore, Z_C has high power for the left-sided alternative indicating that there is significant mixed correspondence in the NN structure. The power estimates for these tests increase as r decreases.

4.2.2. Empirical Power Analysis for Three Classes

For the three classes, we consider two cases. We generate $n_1 = n_2 = n_3 = 40$ points from classes X, Y and Z.

Case 1: For this class of alternatives, we generate X and Y points as in Case III of power analysis for two classes of Section 4.2.1, and Z points as Y points in Case I of Section 4.2.1. We consider the following two alternatives:

$$H_1^1 : s = 1/6, \sigma = 1/5, \quad \text{and } H_1^2 : s = 1/4, \sigma = 2/15. \tag{18}$$

The classes 1 and 2 (i.e., X and Y) are segregated with shifted supports and class 3 is clustered around $(1/2, 1/2)$. Furthermore, by construction a higher level of niche specificity for Z points exists around the center of the unit square compared to X and Y points, which in turn implies segregation of Z points from X and Y points as well.

The empirical power estimates under the alternatives, H_1^1 and H_1^2, with $n_1 = n_2 = n_3 = 40$ are presented in Table 11, where $\widehat{\beta}_{\mathscr{X}_D}$ is power estimate for Dixon's overall test of segregation; $\widehat{\beta}_{Z_{33}}$ is for the cell-specific test for cell $(3,3)$ (for segregation), the other notation is as in Table 9. Under the case 1 alternatives, the power estimates increase as σ decreases and s increases. In particular, Dixon's overall test and the self column test, $\mathscr{X}_C$ have high power estimates, and the right-sided cell-specific tests for cells $(1,1)$, $(2,2)$ and $(3,3)$ have high power, which indicate segregation of each class from the others. Also, the right-sided Z test for the sum of the self column, Z_C, has high power, implying self correspondence is operating as well. Notice that the Z_C has the highest power estimates.

Case 2: For this class of alternatives, we again generate X and Y points as in Case III of Section 4.2.1, and Z points as Y points in Case V of Section 4.2.1. We consider the following two alternatives:

$$H_2^1 : s = 1/6, r = 1/7, \quad \text{and } H_2^2 : s = 1/4, r = 1/10. \tag{19}$$

The classes 1 and 2 (i.e., X and Y) are segregated with shifted supports and class 3 is clustered around X and Y points.

Table 11: *The power estimates under the case 1 and 2 alternatives with $N_{mc} = 10000$, $n_1 = n_2 = n_3 = 40$ at $\alpha = .05$. $\widehat{\beta}_{\mathscr{X}_D}$ is the power estimate for Dixon's overall segregation test, $\widehat{\beta}_{Z_{33}}$ is for the cell-specific test for cell $(3,3)$. The other notation and superscripting for "$<$" and "$>$" are as in Table 9.*

Power estimates under						
case 1 alternatives						
	$\widehat{\beta}_{\mathscr{X}_D}$	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{>}_{Z_C}$	$\widehat{\beta}^{>}_{Z_{11}}$	$\widehat{\beta}^{>}_{Z_{22}}$	$\widehat{\beta}^{>}_{Z_{33}}$
H_1^1	.3297	.4453	.6453	.5186	.5254	.1951
H_1^2	.9527	.9809	.9958	.9606	.9570	.6985
case 2 alternatives						
	$\widehat{\beta}_{\mathscr{X}_D}$	$\widehat{\beta}_{\mathscr{X}_C}$	$\widehat{\beta}^{>}_{Z_C}$	$\widehat{\beta}^{>}_{Z_{11}}$	$\widehat{\beta}^{>}_{Z_{22}}$	$\widehat{\beta}^{<}_{Z_{33}}$
H_2^1	.2620	.2289	.0539	.1877	.1834	.2954
H_2^2	.6514	.4818	.1032	.3488	.3433	.4099

The empirical power estimates under the alternatives, H_2^1 and H_2^2, with $n_1 = n_2 = n_3 = 40$ are presented in Table 11. Under the case 2 alternatives, the power estimates increase as r decreases and s increases. In particular, Dixon's overall test and the self column test, $\mathscr{X}_C$ have high power, and the right-sided cell-specific tests for cells $(1,1)$ and $(2,2)$, and left-sided test for cell $(3,3)$ have high power, which indicate segregation of X and Y points from other classes, and lack of segregation of Z points from X and Y points which might be association of Z points with one or both of classes X and Y. To determine the specifics one needs to check the off-diagonal cell specific tests in row 3 of the corresponding NNCT. Also, the right-sided Z test for the sum of the self column is mildly significant, implying a mild level of self correspondence is operating as well. Notice also that the Z_C has the lowest power estimates.

In alternative cases I-V the classes are either both segregated or associated (i.e., the direction of the deviation in each diagonal cell is same for both classes). Hence this cumulative effect is better captured by Z_C which has the highest power estimates under all these cases. Similarly, in alternative case 1, by construction, each class is segregated from others (although the type and level of segregation is different for class Z compared to X and Y), hence the direction of the deviation in the self column in CCT (i.e., diagonal cells in NNCT) is same for all classes, thereby rendering Z_C the most powerful test again. However, in alternative case 2, while X and Y are segregated, Z is associated with both classes. Hence direction of deviation in the self column cells are positive for X and Y and negative for Z. So this sign difference tends to nullify the deviations from the null case, which causes Z_C have the lowest power estimates.

5. Real-Life Example Data Set

To illustrate the methodology, we use an example data set with 6 classes: the Lansing Woods data which is available in the spatstat package in R (Baddeley and Turner, 2005).

The Lansing Woods data contains locations of trees (in feet (ft)) and botanical classification of trees according to their species in a 924 ft $\times$ 924 ft (19.6 acre) plot in Lansing Woods, Clinton County, Michigan, USA (Gerrard, 1969). The data set comprises of 2251 trees together with their species as hickories, maples, red oaks, white oaks, black oaks and miscellaneous trees. In our analysis, we consider each species as a class and miscellaneous trees as another class. The scatterplot of these tree locations are presented in Figure 2.

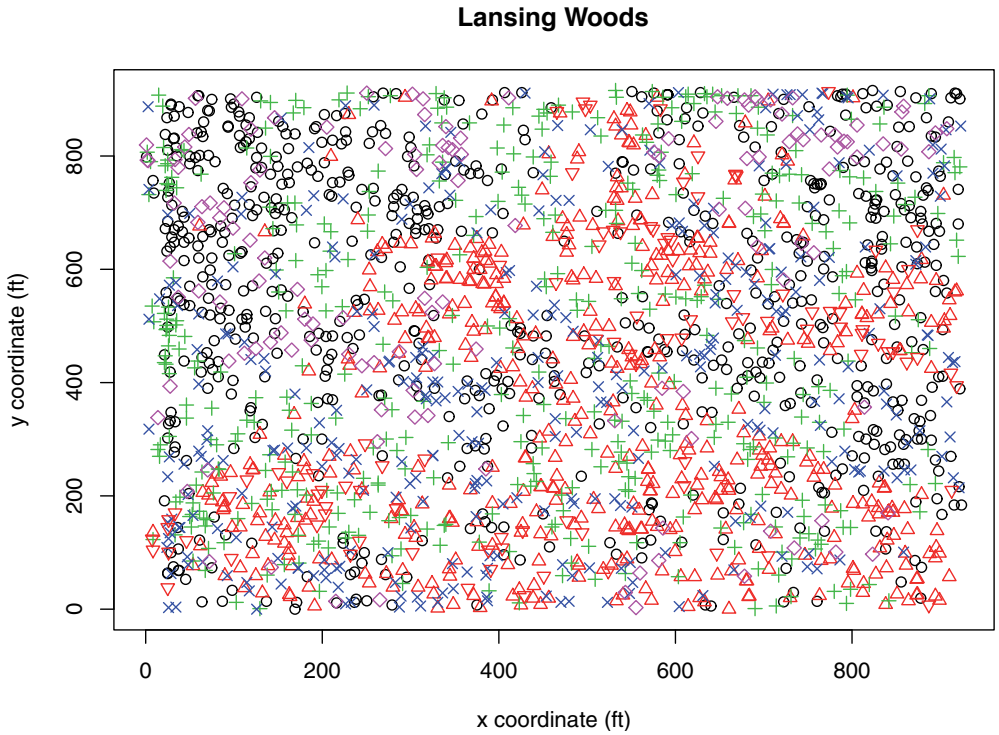


Figure 2: *The scatterplot of the locations of hickories (circles ○), maples (triangles △), white oaks (pluses +), red oaks (crosses ×), black oaks (diamond shapes ⋄), and other species (inverse triangles ▽) in the Lansing Woods, Clinton County, Michigan, USA.*

The CCT for this data is presented in Table 12. Notice that some cell counts in the contingency tables are not integers, since there are ties in the NN relations. For self correspondence, the abundance proportions for the species is hickories:maples:whit oaks:red oaks:black oaks:other $\approx 6 : 70 : 4.90 : 4.27 : 3.30 : 1.29 : 1.00$ and the proportions of the entries in the self column is $\approx 14.14 : 9.70 : 5.50 : 4.20 : 1.08 : 1.00$, which seems to be much different than the abundance proportions, suggesting significant presence of self correspondence.

Table 12: *The CCT for the Lansing Woods data.*

		pair type		
		self	mixed	total
base species	hickory	353.5	349.5	703
	maple	242.5	271.5	514
	white oak	137.5	310.5	448
	red oak	105	241	346
	black oak	27	108	135
	other	25	80	105
	total	890.5	1360.5	2251

Table 13: *The test statistics and the p-values for Lansing Woods data. Z_{ii} are cell-specific tests for cells (i,i) for $i = 1,2,\ldots,6$, $\mathscr{X}_D$ is Dixon's overall test of segregation. Z_C, and $\mathscr{X}_C$ are as defined in the text; TS stands for the test statistic, p_{asy}, p_{mc}, and p_{rand} stand for the p-values based on the asymptotic approximation, Monte Carlo simulation, and randomization of the tests, respectively.*

Test statistics and p-values for Lansing Woods data									
	$\mathscr{X}_D$	$\mathscr{X}_C$	$Z_S^>$	$Z_{11}^>$	$Z_{22}^>$	$Z_{33}^>$	$Z_{44}^>$	$Z_{55}^>$	$Z_{66}^>$
TS	376.8609	325.9750	16.4759	9.4622	11.0934	4.7895	6.3717	5.5085	7.4514
p_{asy}	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$
p_{rand}	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$
p_{mc}	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$	$<.0001$

We compute $Q = 1560$ and $R = 1400$. Again the more appropriate null hypothesis is the CSR independence pattern, since the locations of the tree species can be viewed a priori resulting from different processes (Goreaud and Pélissier, 2003). We present the test statistics and the associated p-values in Table 13, where in this table p_{asy} stands for the p-value based on the asymptotic approximation (i.e., asymptotic critical value), p_{rand} is based on Monte Carlo randomization of the labels on the given locations of the trees 10000 times and p_{mc} is the p-value based on 10000 Monte Carlo replication of the CSR independence pattern in the region plotted in Figure 2. Notice that p_{asy}, p_{rand} and p_{mc} are similar and highly significant for all tests. The cell-specific tests are all significant for the right-sided alternative, and the χ^2 test for the self column, $\mathscr{X}_C$, is significant, implying significant self correspondence for these species, and hence significant segregation of the species (from each other). Similarly, Z_C is significant confirming significant self correspondence for all species combined.

6. Discussion and Conclusions

In this article, we introduce the correspondence in the NN structure pattern for multiple classes/species and tests for it based on a contingency table called correspondence contingency table (CCT) which can also be derived from the associated nearest neighbour

contingency table (NNCT). These tests are a χ^2 test of correspondence for the first column of CCT (called self-column), $\mathscr{X}_C$, and a Z test for the sum of the self column of CCT, Z_C. We show that in the two class case, the CCT and the NNCT contain the same information (but in different order in their entries), and the corresponding quadratic test for the self column, $\mathscr{X}_C$ and Dixon's overall test of segregation, $\mathscr{X}_D$, are equivalent. For more than two classes, these tests are different and hence provide different information. On the other hand, regardless of the number of classes, Z_C is different from $\mathscr{X}_C$ and $\mathscr{X}_D$ (i.e., Z_C provides new information not provided by the segregation tests for two or more classes) whereas Z_{S_i} and Z_{ii} are identical (i.e., they always give the same information).

For $k \geq 2$ classes, NNCT is of dimension $k \times k$ and the corresponding CCT is of dimension $k \times 2$, where the entries in the first column (i.e., self column) are the diagonal entries of the NNCT and each entry in the second column (i.e., mixed column) of CCT is the sum of off-diagonal entries at each row of NNCT. Overall segregation test based on NNCT measures any deviation in the entries of the NNCT and a cell-specific test based on NNCT measures the deviation in the corresponding entry of the NNCT (Dixon, 1994, 2002b). On the other hand, the tests based on the CCT are a χ^2 test for the self column and a Z test for the sum of the self column. The former test is based on deviations of the frequencies of the self NN pairs, and the latter is based on the sum of these frequencies. Both tests might indicate presence of self or mixed correspondence which can not be tested directly in the NNCT, hence the need to introduce CCT.

We show that $\mathscr{X}_C$ provides information on the overall deviations jointly in the self column (or in the mixed column) in CCT, Z_C provides information on the abundance of self pairs when all classes are combined. Hence to determine the level and type of correspondence as self or mixed, $\mathscr{X}_C$ should be employed together with the cell-specific tests Z_{ii} (see Equation (5)) so that when $\mathscr{X}_C$ is significant cell-specific tests will provide the direction and significance of the deviations for each diagonal cell in the NNCT (or each cell in the self-column of CCT). If they are all or mostly in the positive (resp. negative) direction, the pattern would be segregation (resp. lack of segregation) for the classes corresponding to the positive (resp. negative) significant Z_{ii} values and self (resp. mixed) correspondence for all classes combined. On the other hand, for Z_C we do not need to confer to the cell-specific tests, as it by itself is sufficient to indicate that the correspondence is of type self or mixed. Another advantage of Z_C is that it is more robust to differences in relative abundances of the classes (i.e., to the class imbalance problem).

Among the tests considered, Z_C is more powerful if all or most classes are segregated. The same holds if all or most class pairs are associated. But if the pattern is mixed (i.e., some classes are segregated while some pairs are associated) the deviations in the self column tend to cancel each other in the sum, rendering Z_C perform rather poorly. In such a case, $\mathscr{X}_C$ (together with the cell-specific tests) provide a more accurate picture of the patterns in the data and are more powerful.

Based on our simulations and example data sets, we recommend to perform both of the tests Z_C and $\mathscr{X}_C$, and if any of them is significant, then the cell-specific tests can be performed (to determine segregation or lack of it at the class/species level). When the cell counts in the self column of CCT are all larger than 10, it is safe to employ $\mathscr{X}_C$ with the asymptotic approximation, and if some cell count is 5 or less, it is better to use Monte Carlo randomized version of the test. If some cell counts are between 5 and 10, both versions (i.e., asymptotic approximation and Monte Carlo randomization) can be used to reach more reliable conclusions. Since Z_C is the sum of the self column, the cell counts are not that relevant as long as column sum is 20 or larger (even 10 or larger seems to work in practice). We recommend randomization version of Z_C if column sum is 10 or less, and for sum between 10 and 20, one can employ both asymptotic approximation and randomization versions for more reliable conclusions.

Throughout this article, we assume the total sample size and class sizes are all fixed. If it is desired to have the sample size be a random variable, we may consider a spatial Poisson point process on the region of interest instead of the binomial process. In fact, this case is also a realistic situation for data collection schemes in plant ecology. That is, in the region of interest, one can examine each subject, determine its species and that of its NN. In this framework, all margins of the NNCT and CCT would be random. The effect of such randomness on the behavior (e.g., distribution), size and power performance of the tests is a topic of prospective research. For the cases where CSR independence is the appropriate benchmark (see Section 3.1), this framework might be more realistic, but for the cases where RL is the appropriate benchmark, then our approach in this article is more realistic.

We have discussed the patterns of segregation and (self and mixed) correspondence mostly in the context of plant ecology. However, the patterns and the associated tests can be applied in other contexts as well. For example, one can apply them in an epidemiological or a social context by using the residences of people as their location. In the epidemiological context, the question of interest could be the distribution (i.e., clustering or lack of it) of a disease. In disease clustering, significant segregation of disease cases can have further implications (e.g., one can then search for the reasons of such clustering which can help in controlling the spread of the disease or curing the diseased people). In the social context of racial distribution of residences, segregation of any particular race would imply their clustering in certain neighbourhoods; self correspondence would mean that all racial groups tend to live in clumps or clusters of same race residents (i.e., there is lack of local diversity in the region) On the other hand, mixed correspondence of racial status of residents would imply that the society is diverse at the local level as people of different races live side by side in a mixed neighbourhood and there is no preference of the residents to live by people of the same race.

In the literature, usually NN relationships are based on the distance metrics. For example, in this article, Euclidean distance in $\mathbb{R}^2$ is the only metric used. The NN relations based on dissimilarity measures is an extension of NN relations based on distance metrics. In such an extension, NN of an object, x, refers to the object with the minimum

dissimilarity to x. We assume that the objects (events) lie in a finite or infinite dimensional space satisfying the lack of any inter-dependence which implies lack of self or mixed correspondence in the NN structure. Under RL, the objects' locations are fixed yielding fixed interpoint dissimilarity measures, but the labels are assigned randomly to the objects. Although our correspondence tests are constructed assuming data are in $\mathbb{R}^2$, the extension to higher dimensions is straightforward.

Acknowledgments

I would like to thank an anonymous associate editor and two referees, whose constructive comments and suggestions greatly improved the presentation and flow of the paper. This research was supported by the research agency TUBITAK via Project # 111T767 and by the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD.

References

Baddeley, A. J. and R. Turner (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12, 1–42.

Ceyhan, E. (2008a). Overall and pairwise segregation tests based on nearest neighbor contingency tables. *Computational Statistics & Data Analysis*, 53, 2786–2808.

Ceyhan, E. (2008b). QR-adjustment for clustering tests based on nearest neighbor contingency tables. arXiv:0807.4231v1 [stat.ME]. Technical Report # KU-EC-08-5, Koç University, Istanbul, Turkey.

Ceyhan, E. (2010). On the use of nearest neighbor contingency tables for testing spatial segregation. *Environmental and Ecological Statistics*, 17, 247–282.

Ceyhan, E. (2014). Nearest neighbor methods for testing reflexivity and species-correspondence. arXiv: 1405.3689 [stat.ME]. Technical Report # KU-EC-14-1, Koç University, Istanbul, Turkey.

Clark, P. J. and F. C. Evans (1954). Distance to nearest neighbor as a measure of spatial relations in population ecology. *Ecology*, 35, 445–453.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman & Hall / CRC, Boca Raton, FL.

Cox, T. F. (1981). Reflexive nearest neighbours. *Biometrics*, 37, 367–369.

Cuzick, J. and R. Edwards (1990). Spatial clustering for inhomogeneous populations (with discussion). *Journal of the Royal Statistical Society, Series B*, 52, 73–104.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns, 2nd edition*. Hodder Arnold Publishers, London.

Dixon, P. M. (1994). Testing spatial segregation using a nearest-neighbor contingency table. *Ecology*, 75, 1940–1948.

Dixon, P. M. (2002a). Nearest-neighbor contingency table analysis of spatial segregation for several species. *Ecoscience*, 9, 142–151.

Dixon, P. M. (2002b). Nearest neighbor methods. *Encyclopedia of Environmetrics, edited by Abdel H. El-Shaarawi and Walter W. Piegorsch, John Wiley & Sons Ltd., NY*, 3, 1370–1383.

Fargione, J. and D. Tilman (2005). Niche differences in phenology and rooting depth promote coexistence with a dominant C_4 bunchgrass. *Oecologia*, 143, 598–606.

Gerrard, D. J. (1969). *Competition quotient: a new measure of the competition affecting individual forest trees. Research Bulletin, Agricultural Experiment Station*, Volume 20. Michigan State University.

Goreaud, F. and R. Pélissier (2003). Avoiding misinterpretation of biotic interactions with the intertype K_{12}-function: population independence vs. random labelling hypotheses. *Journal of Vegetation Science*, 14, 681–692.

Kulldorff, M. (2006). Tests for spatial randomness adjusted for an inhomogeneity: A general framework. *Journal of the American Statistical Association*, 101, 1289–1305.

Pielou, E. C. (1961). Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships. *Journal of Ecology*, 49, 255–269.

Primack, R. (1998). *Essentials of Conservation Biology*. Sunderland: Sinauer Associates.

Ranker, T. A. and C. H. E. Haufler (2008). *Biology and Evolution of Ferns and Lycophytes*. Cambridge University Press, Cambridge, UK.

Ripley, B. D. (2004). *Spatial Statistics, 2nd edition*. Wiley-Interscience, New York.

Searle, S. R. (2006). *Matrix Algebra Useful for Statistics*. Wiley-Interscience, New York.

van Lieshout, M. N. M. and A. J. Baddeley (1999). Indices of dependence between types in multivariate point patterns. *Scandinavian Journal of Statistics*, 26, 511–532.

Vichi, M. and G. Saporta (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53, 3194–3208.

Werner, E. E. and J. F. Gilliam (1984). The ontogenetic niche and species interactions in size-structured populations. *Annual Review of Ecology and Systematics*, 15, 393–425.

SORT 42 (2) July-December 2018, 159-182 DOI: 10.2436/20.8080.02.73

Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

Ramón Ferri-García and María del Mar Rueda*

Abstract

One of the main sources of inaccuracy in modern survey techniques, such as online and smartphone surveys, is the absence of an adequate sampling frame that could provide a probabilistic sampling. This kind of data collection leads to the presence of high amounts of bias in final estimates of the survey, specially if the estimated variables (also known as target variables) have some influence on the decision of the respondent to participate in the survey. Various correction techniques, such as calibration and propensity score adjustment or PSA, can be applied to remove the bias. This study attempts to analyse the efficiency of correction techniques in multiple situations, applying a combination of propensity score adjustment and calibration on both types of variables (correlated and not correlated with the missing data mechanism) and testing the use of a reference survey to get the population totals for calibration variables. The study was performed using a simulation of a fictitious population of potential voters and a real volunteer survey aimed to a population for which a complete census was available. Results showed that PSA combined with calibration results in a bias removal considerably larger when compared with calibration with no prior adjustment. Results also showed that using population totals from the estimates of a reference survey instead of the available population data does not make a difference in estimates accuracy, although it can contribute to slightly increment the variance of the estimator.

MSC: 62D05

Keywords: Online surveys, Smartphone surveys, propensity score adjustment, calibration, simulation

1. Introduction

Traditional surveys are experiencing, along with society, a number of changes which affect their validity and applicability. Several reasons can be cited (e.g., see Couper, 2017, Schonlau et al., 2009) on the decline of participation and completion rates in surveys conducted using traditional modes of contact, such as telephone or face-to–

* *Corresponding author:* mrueda@ugr.es
Department of Statistics and Operations Research, University of Granada, Avda. Fuentenueva s/n 18004 Granada, Spain. rferri@ugr.es, mrueda@ugr.es
Received: April 2018
Accepted: November 2018

face surveys. A review performed by Díaz de Rada (2012) stated that response rates in traditional surveys have been dropping for two decades. The increasing difficulty of contacting households members in face-to-face surveys results in increased costs per interview and therefore non-sampling errors are problematic to deal with in this context; regarding telephone surveys, the rise of mobile phones makes it more difficult for government agencies to keep an adequate sampling frame, in terms of coverage, of landline phones (Pasadas-del-Amo, 2018).

At the same time, the arrival of the internet and mobile phone lines has led to the usage of new survey administration methods, with online surveys and smartphone surveys being the most popular and promising ones to deal with the above mentioned issues in order to contact respondents. Online surveys can be defined, given how they are conducted nowadays as described by Mei and Brown (2017), as surveys completed from computers that respondents can access anytime. Questionnaires might have a conventional structure adapted to the online context (e.g., SurveyMonkey) and might also be provided using online social networks. Smartphone surveys differ in the mode in which they are completed: any survey completed using a mobile device or a tablet can be considered a smartphone survey. Sometimes, the questionnaire might be hosted in an URL, thus it could be considered a browser survey and therefore an online survey. This states a clear divide in the smartphone surveys between those app-based questionnaires or related and those completed using a browser available in the device itself, as the latter do not properly seize the advantages of a mobile device.

The change from the traditional survey to the internet survey has brought important changes and new challenges have arisen (Díaz de Rada and Domínguez, 2015, 2016). These new methods offer substantial advantages against traditional survey techniques, specially in terms of monetary and time costs as they usually do not require any effort by any interviewer and the information collection becomes instantaneous. In addition, online surveys are considered to be more advantageous for information collection; despite the advantages of smartphones such as the audiovisual options and the possibility to retrieve data on certain variables without the need of any extra question in the survey, web surveys take less time to be completed by interviewers, as proved by Couper and Peterson (2017).

Along with the described advantages, some serious concerns often arise when using these new survey methods. As noted in Elliott and Valliant (2017), internet surveys (even when a structured voluntary panel is used) suffer mostly from selection bias, specially from the bias induced by the internet availability and penetration in the general population. This issue will be broadly discussed later. Internet surveys are also affected by nonresponse bias; a meta-analysis conducted by Manfreda et al. (2008) estimated that online surveys are associated with a decrease in response rates between 6% and 15% in comparison to other survey modes. In addition, the use of incentives as a method to improve cooperation have been proved as less efficient in online surveys (Díaz de Rada, 2012). Other important sources of non-sampling errors in online and smartphone surveys are measurement errors; although the social desirability effect is less prone to

appear in online surveys (Heerwegh, 2009), they still suffer from other effects such as technical issues (e.g., poor internet connection may lead to a lack of completion of a survey), or lack of veracity in the responses given, which in the online case has a variety of causes.

Nonresponse bias, as well as measurement errors, have been widely studied in the literature as they have been common issues in traditional survey methods since their initial development. However, selection bias presents some particular characteristics in the new survey methods which require other strategies in order to tackle it. In all cases, online and smartphone surveys are often applied under inadequate sampling conditions; they are generally taken by self-selected respondents which conform a non-probabilistic sampling. Even if an acceptable random sampling is eventually performed, it may be particularly troublesome to establish a reliable sampling frame to meet the probabilistic sampling assumptions (Couper, 2000, Couper and Peterson, 2017). On the other hand, the coverage of such surveys is also limited by the population access to the internet. Although no interview mode is exempt from suffering coverage bias, it happens to be much more important in internet surveys (Couper (2007), according to Schonlau et al. (2009)), as internet access is often associated with sociodemographic variables which could be eventually related to the outcome variables of a certain study. To mention some examples, data from the Pew Research Center (2017) reveal that in 2016 while 99% of U.S. adults between 18 and 29 years old could be considered internet users, only a 64% of those above 65 years of age fell into the same group. In the case of Spain, the generation gap is wider according to the National Institute of Statistics (2017a); while the internet penetration rate is above 90% for all age groups below 54 years of age, in citizens between 65 and 74 years old penetration rate is 43.7%.

It is obvious that such a problem can be responsible for a large increase in the bias of the final results. Therefore, developing methods to deal with the lack of representativity has become a priority. To date, the more relevant methods are considered to be calibration techniques and propensity score adjustment (PSA). Calibration weighting using auxiliary information (Deville and Särndal, 1992) has been established as the main technique to deal with problematic sampling frames, but its efficacy can decrease when the self-selection procedure is tied, directly or not, to the target variables (Bethlehem, 2010). Calibration for coverage issues has also been studied using the superpopulation model approach through general regression (GREG) weights (Dever, Rafferty and Valliant, 2008); even though it successfully address both nonresponse and noncoverage in online surveys, it requires an structured sampling design, something that does not apply to volunteer surveys. When calibration is ineffective, PSA can be a proper substitute if it is feasible to use a probabilistic sample on the same target population, on which a subset of variables measured on the non-probabilistic sample have been measured on the probabilistic sample as well. Research findings have shown that PSA successfully removes bias in some situations, but at the cost of increasing the variance of the estimates (Lee, 2006, Lee and Valliant, 2009). The efficacy of bias removal by PSA is strongly dependent on using covariates related to the actual propensity to participate

and the target variables (Schonlau and Couper, 2017), and its sole application without any further adjustment can lead to biased estimates (Valliant and Dever, 2011). The aim of this study was to examine the behaviour of the estimators when both techniques, PSA and calibration, are applied, in comparison to the situations where only calibration is performed or where no weighting technique is applied at all. Given that, for most situations, auxiliary information can be troublesome to find, calibration is tested using known population totals and using population estimates coming from the reference (probabilistic) sample that it is supposed to be available. Under the initial hypothesis of the study, the combined weighting of PSA in a first step and calibration in a second one would outperform the estimates obtained with calibration weighting only in terms of bias reduction, although the estimators will have a higher variance as the reference sample size gets smaller in comparison to the convenience (non-probabilistic) sample size.

2. Methodology

2.1. Calibration weighting

Surveys often have a coverage error associated to them, in the sense of being made using a sampling frame that does not cover the entire population to which survey results are to be extrapolated. This coverage error, which can be the result of several irregularities, can be controlled by the use of reweighting or calibration techniques. Calibration was defined by Särndal (2007) as the combination of three items: "a) a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s), b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units, c) an objective to obtain nearly design unbiased estimates as long as non-response and other non-sampling errors are absent".

Calibration theory can be explained as follows (Deville and Särndal, 1992): let y be the variable of interest in the survey estimation and s the sample collected in the survey, with each element k in the sample having an associated probability of selection, $\pi_k = 1/d_k$. Without any auxiliary information, the population total of y, Y, is estimated in a non-biased way with the Horvitz-Thompson estimator:

$$\hat{Y}_{HT} = \sum_{k \in a} d_k y_k \tag{1}$$

Let $\mathbf{x}$ be an auxiliary vector associated to y, with population total assumed to be known $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$. The calibration estimation of Y consists in the obtaining of a new weights vector w_k for $k \in s$ which modifies as little as possible the original sample weights, d_k, which have the desirable property of producing unbiased estimations, respecting at the same time the calibration equations:

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}. \tag{2}$$

Given a distance $G(w_k, d_k)$, the calibration process consists on finding the solution to the minimization problem

$$\min_{w_k} E\{\sum_{k \in s} G(w_k, d_k)\} \tag{3}$$

while respecting the calibration equation (2). Several distances were defined in Deville and Särndal (1992), the linear distance being one of the most commonly used (Rueda et al., 2010, Martínez et al., 2010). This distance is calculated by:

$$\sum_{k \in S} \frac{(w_k - d_k)^2}{q_k d_k} \tag{4}$$

q_k are positive weights that are usually assumed as uniform (i. e. $1/q_k = 1$), although unequal weights $1/q_k$ are sometimes used. The problem now concerns finding the minimum of (4) subject to (2), leading to the calibrated weight:

$$w_k = d_k(1 + q_k \mathbf{x}_k' \lambda) \tag{5}$$

where the vector of multipliers, λ, is calculated as:

$$\lambda = T_s^{-1}(\mathbf{X} - \sum_s \mathbf{x}_k d_k) \tag{6}$$

T_s, whose inverse is assumed to exist, is the equivalent of:

$$T_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k' \tag{7}$$

The resulting estimator of Y is the general regression estimator (Cassel, Särndal and Wretman, 1976)

$$Y = \sum_s w_k y_k = \sum_s y_k d_k + (\mathbf{X} - \sum_s \mathbf{x}_k d_k)' \hat{B}_s \tag{8}$$

where $\hat{B}_s$ is

$$\hat{B}_s = T_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k \tag{9}$$

In general, the resulting estimator for Y is biased, but it is assumed to be asymptotically unbiased as the new weights w_k would approach the sampling weights d_k.

2.2. Propensity score adjustment (PSA)

The propensity score adjustment method was originally developed by Rosenbaum and Rubin (1983) which sought to reduce the bias due to treatment and control assignment in non-randomized studies. The main idea of the adjustment is to balance the differences between groups in non-randomized designs with the computation of a score whose distribution is the same for all groups. The proposed score for a given unit is equivalent to its probability of being in the treatment group, which can be estimated using a regression model. Although the implications of this approach in survey nonresponse were considered shortly after Rubin (1986), according to Little and Rubin (2002), it was not proposed for online surveys until Harris Interactive took it into account in their internet research (Taylor, 2000, 2001). To a lesser extent, these first attempts added one element to the requirements for performing PSA: a reference survey. The concept of reference survey was extended in further studies (see Lee, 2006).

When treating an online survey, it is expected that the sampling was conducted in a non-probabilistic manner or even not conducted at all, with the survey being completed by volunteer respondents. It is feasible to consider that the decision to take part on the survey depends on a probability which, depending on the respondent characteristics, might be higher or lower. In this case, a reference survey can be very helpful to determine this probability. A reference survey is conducted on the same target population than the online survey, with the main difference that the former has a better coverage and higher response rates than the latter, thus it is adequate to represent the behaviour that the target population should have when a probabilistic survey is performed on it.

Once data is collected from both surveys, the propensity for an individual to take part on the volunteer (non-probabilistic) survey is obtained by binning the data together and training a logistic regression model on the dichotomous variable, z, which measures whether the respondent took part in the volunteer survey or in the reference survey. The model uses covariates, $\mathbf{x}$, that have been measured in both surveys, thus the formula to compute the propensity of taking part in the volunteer survey, π, can be displayed as

$$\pi(x) = \frac{1}{e^{-(\gamma^{\mathrm{T}}\mathbf{x}_k)} + 1} \tag{10}$$

for some vector γ, as a function of the model covariates.

We denote by s_R the reference sample and by s_V the volunteer sample. Following the approach described in Lee and Valliant (2009) which will be used in this study, propensity scores are divided in g classes, with $g = 5$ as the conventional choice following Cochran (1968), where all units may have the same propensity score or at least be in a very narrow range. For each class, an adjustment factor is calculated as stated in (11):

$$f_g = \frac{\sum_{k \in s_{Rg}} d_{Rk} / \sum_{k \in s_R} d_{Rk}}{\sum_{k \in s_{Vg}} d_{Vk} / \sum_{k \in s_V} d_{Vk}} \tag{11}$$

where s_{Rg} is the set of individuals in the reference sample that are in the gth class of propensity scores, and d_{Rk} is the original design weight of the k individual in the reference sample, s_{Vg} is the set of individuals in the volunteer sample that are in the gth class of propensity scores, and d_{Vk} is the original design weight of the k individual in the volunteer sample. Finally, the adjusted weights d^* are the product of the original weights and the adjustment factor; following the same notation, the adjusted weight for individual k in s_{Vg} (i. e. the individual k of the gth propensity class in the volunteer sample) is computed as indicated in (12). These weights are equivalent to the weights used for the Horvitz-Thompson (H-T) estimator.

$$d_k^* = f_g d_{Vk} = \frac{\sum_{k\in s_{Rg}} d_{Rk} / \sum_{k\in s_R} d_{Rk}}{\sum_{k\in s_{Vg}} d_{Vk} / \sum_{k\in s_V} d_{Vk}} d_{Vk} \tag{12}$$

Alternatively, the approach proposed by Schonlau and Couper (2017) can be used to obtain weights for a Hajek-type estimator using propensity scores. This approach has the particularity of adjusting to the population of the probabilistic sample, rather than the combined population of the two samples. Weights are defined as the inverse propensity scores, as indicated in (13)

$$w_i = \frac{1-\hat{\pi}(\mathbf{x}_k)}{\hat{\pi}(\mathbf{x}_k)} \tag{13}$$

where $\hat{\pi}(\mathbf{x}_k)$ is the estimated response propensity for the individual k of the volunteer sample as predicted by logistic regression with covariates $\mathbf{x}$.

3. Simulation study

3.1. Data description

To explore the effectivity of PSA with further calibration compared to calibration alone, a fictitious population was simulated in order to analyse and establish conclusions for the behaviour of these techniques when applied in real situations. The simulation was based on the study presented in Bethlehem (2010), introducing several changes to extend the spectrum of possible cases in which adjustment methods can be used. In the proposed simulation study, a survey would be conducted to examine a population's voting intention. The population had a fixed size of $N = 50000$, and six variables were included in the study: age, nationality (native/non-native), gender, education (primary/secondary/tertiary), access to the internet (yes/no), and party to which they intended to vote, with four possible options: Party 1, Party 2, Party 3 and Abstention. The distribution of the variables and the relationships between them were fixed as follows:

Table 1: *Probability of each education level as the highest achieved by the fictitious individual, by age groups.*

Education level/Age group	$<$ 35 years old	35-65 years old	$>$ 65 years old
Primary education	0.35	0.45	0.8
Secondary education	0.2	0.25	0.1
Tertiary education	0.45	0.3	0.1

Table 2: *Probability of access to the internet by a given individual, by age groups and nationality.*

Nationality/Age group	$<$ 35 years old	35-65 years old	$>$ 65 years old
Native	0.9	0.7	0.5
Non-native	0.2	0.1	0.0

- Age followed a beta distribution with $\alpha = 2$ and $\beta = 3$ to make it similar to the Spanish population pyramid (National Institute of Statistics, 2017b), and it ranged from 18 to 100 years old.
- Probability of being non-native depended on the age, which was divided in three classes ($<$ 35, 35-65, and $>$65 years old) and individuals on each had a probability of 0.15, 0.1 and 0.025 respectively of being non-native. This probability is similar to the nationality distribution by ages in Spain (National Institute of Statistics, 2016).
- Probability of being a woman was fixed at 0.5 for everyone, except for individuals above 75 years old, whose probability of being a woman was 0.65, as women in Spain tend to have a greater representation in older ages (National Institute of Statistics, 2017b).
- Probabilities of having a specific education level were fixed to resemble as much as possible the Spanish adult population (National Institute of Statistics, 2017c). These probabilities can be consulted in Table 1.
- Access to the internet was made dependent of two variables: age and nationality. This time the probabilities assignment was not based in real data, in order to capture more patterns in the experiment. Probability of access by age groups and nationalities can be consulted in Table 2.
- Probability of voting for each party depended on the party itself. The following relationships were established to make sure all kinds of missing data mechanisms would be represented in the analysis:
  - Voting for Party 1 depended on the gender of the individual; women had a probability of 0.2 to vote for this party while men had a 0.0 probability. Gender is not related to internet access (which is the responsible for non-response) thus the missing data mechanism could be considered as MCAR (Missing Completely At Random).

– Voting for Party 2 depended on the age of the individual; voting probability was 0.0 for people younger than 35 years old, 0.4 for people between 35 and 65 years old, and 0.6 for people older than 65 years old. Given that age, which is an auxiliar variable, is related to internet access, the missing data mechanism was MAR (Missing At Random).
– Voting for Party 3 depended on the access to the internet and the age; people with no access to the internet had a 0.1 probability, no matter how old they were, while people with access had a 0.6, 0.4 and 0.2 probability for each respective age group. In this case, the target variable is directly related to the non-response mechanism, configuring a NMAR (Not Missing At Random) situation.

3.2. Results

To estimate the bias for every possible situation, several configurations of sample sizes for the volunteer sample were considered, letting it vary between 500 and 10,000 individuals. On the other hand, the reference sample size was fixed in 500 individuals for all the experiments. For each volunteer sample size, 1,000 simulations were computed for the results on estimated percent of vote for each of the parties, using the following methods:

- Non-adjusted (unweighted) estimates from the volunteer sample.
- Calibrating the volunteer sample with population totals or estimated population totals (from the reference sample).
- Reweighting with PSA and applying those weights directly to the sample with no further adjustments.
- Reweighting with PSA and calibrating those weights with population totals or estimated population totals (from the reference sample).

Propensity scores were calculated using both approaches presented in Section 2.2 (with $g = 5$ for stratification in the Horvitz-Thompson estimator weights computation). Variables used for PSA and calibration were assigned in four different situations with the following combinations:

- Situation 1: age and education as PSA covariates, gender as calibration variable.
- Situation 2: age and education as PSA covariates, nationality as calibration variable.
- Situation 3: age and nationality as PSA covariates, education as calibration variable.
- Situation 4: age and nationality as PSA covariates, gender as calibration variable.

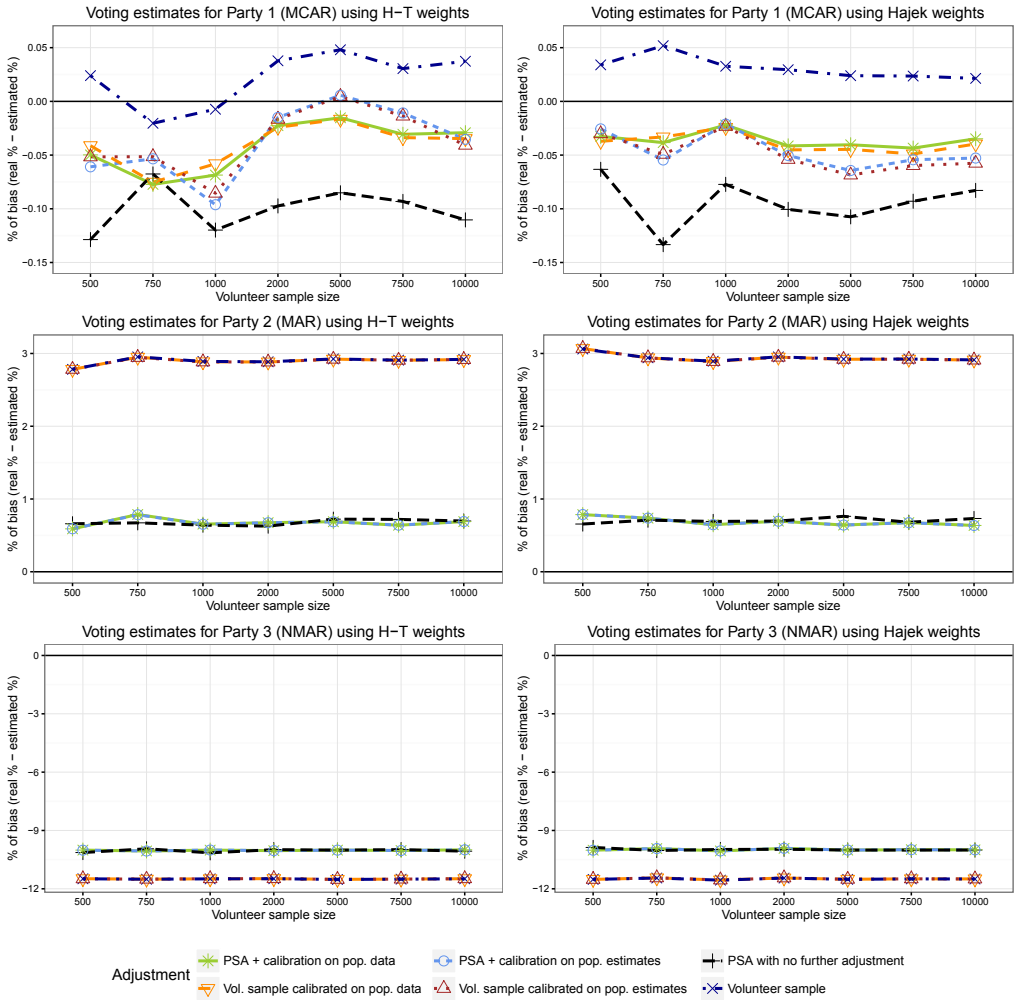


Figure 1: *Bias of each method in voting intention estimations by party in Situation 1.*

For each method and situation, the bias, as a result of the difference between real vote % and estimated vote %, was calculated, as well as the standard deviation of the voting estimation for the 1000 simulations. Figures 1 and 2 summarize results for Situation 1.

Results showed that the difference in bias when the missing data mechanism was completely random is negligible; however, when data was MAR or NMAR, using PSA (regardless of doing calibration afterwards or not) resulted in a reduction in the amount of bias, although this reduction was much higher when data is MAR. It is worth mentioning that these statements could be extended to all the studied sample size situations.

In terms of standard deviations, which give a measure of the variance of the estimator for each method, it can be observed that methods involving PSA resulted in an increase in variance in comparison to methods involving calibration only. However, it is important to point out that the use of estimates of population totals did not increase

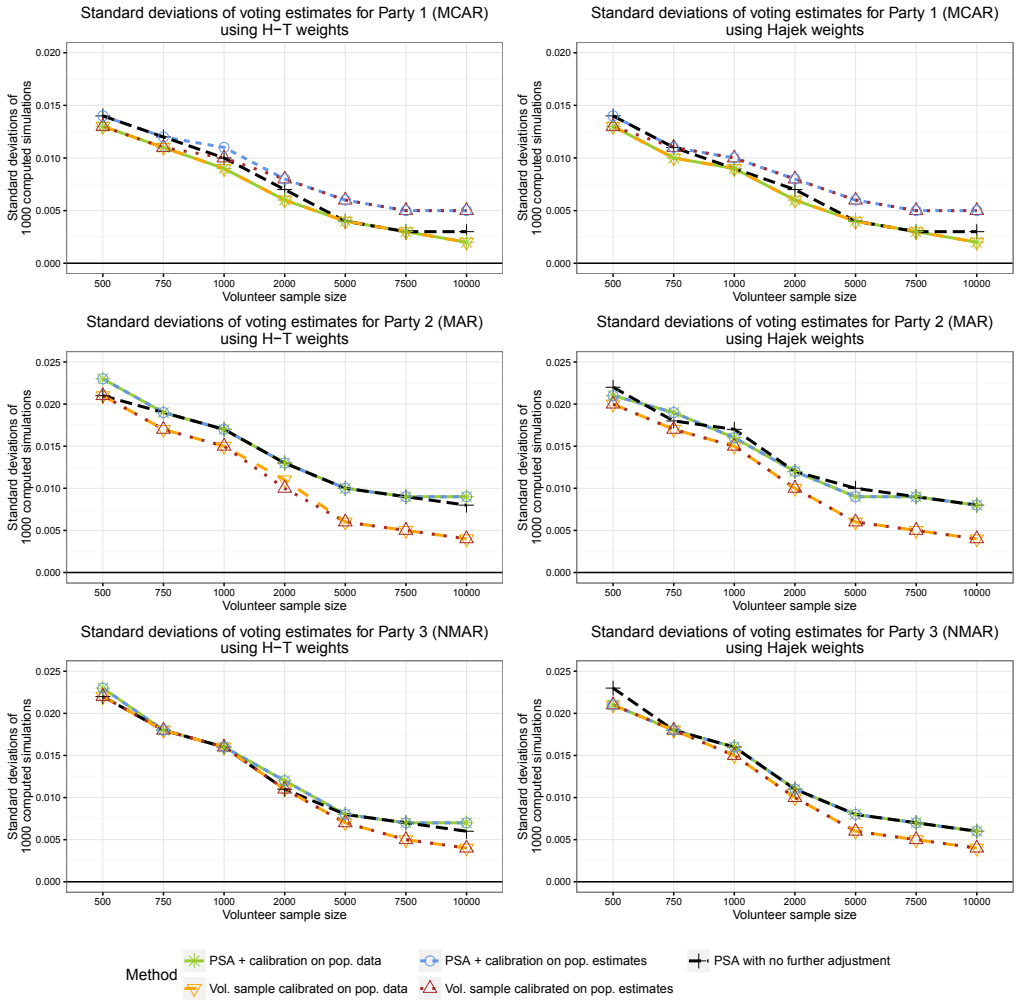


Figure 2: *Standard deviation of voting intention estimations by party provided by each method in Situation 1.*

variance of the survey estimates in MAR and NMAR cases. For the MCAR case, methods involving estimates of population totals resulted overall in greater variance of the estimators.

It is worth mentioning that using Horvitz-Thompson weights or Hajek weights after the computation of the PSA scores made almost no difference in final results in terms of bias reduction or estimators' variance. The very slight differences that could be observed between results may be attributed to the randomness of the experiment rather to an actual effect of the type of weighting.

Figures 3 and 4 summarize results for Situation 2. Bias reduction kept its consistence between weighting methods (Horvitz-Thompson and Hajek), but some differences were found in reference to Situation 1. The only difference between them

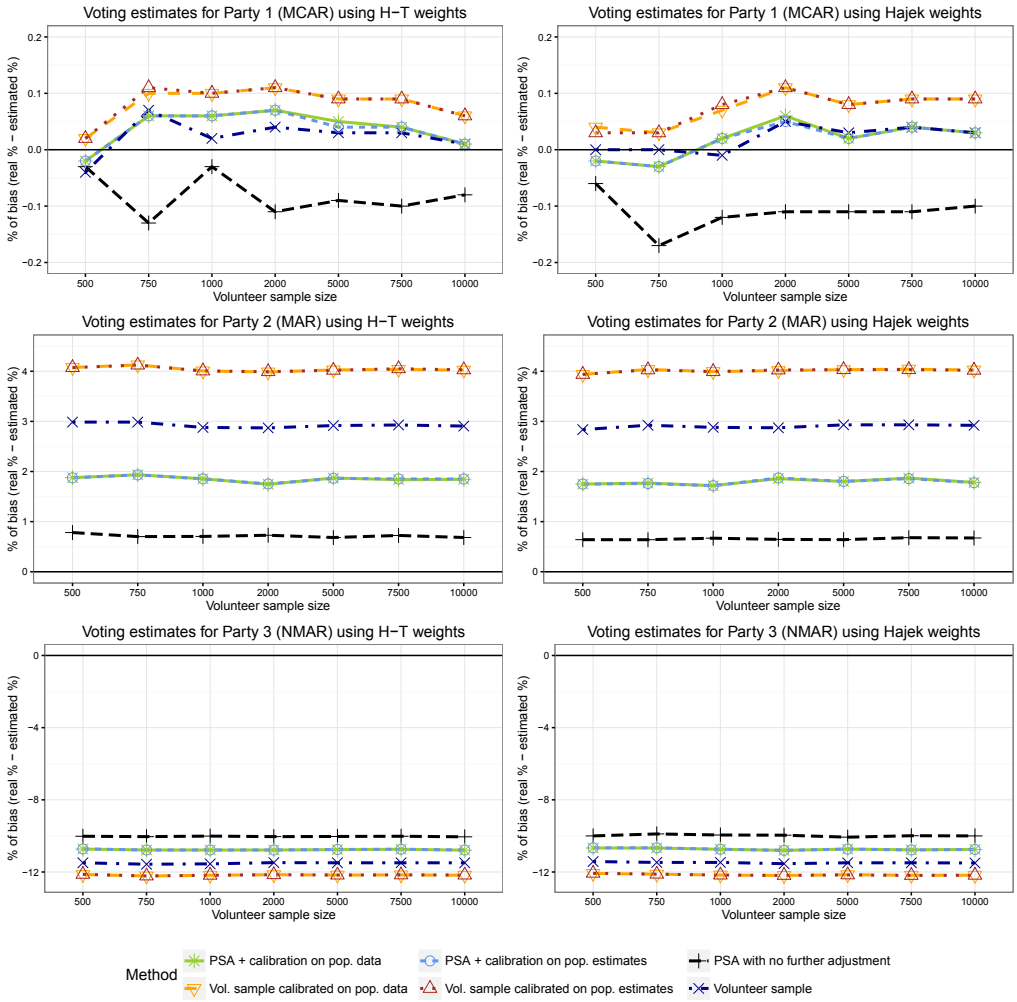


Figure 3: *Bias of each method in voting intention estimations by party in Situation 2.*

was the calibration variable used (nationality instead of gender), but it turned out to be a critical choice. As it can be seen in Figure 3, the application of calibration in Situation 2 resulted in an increase of bias on the estimates, while PSA with no further adjustment produced the same bias reduction than the registered in Situation 1. Estimates involving calibration also had a higher variance, as it can be observed in Figure 4.

Figures 5 and 6 summarize results for Situation 3. In this case, there is a difference in bias reduction motivated by the weighting method used. It is noticeable that Hajek-type estimates are less biased than Horvitz-Thompson-type estimates in the MCAR and MAR cases. It is also worth mentioning that PSA with calibration removed more bias than PSA with no adjustment in the MAR case using Horvitz-Thompson weights. On the contrary, in the NMAR case Horvitz-Thompson-type estimates are less biased than Hajek-type estimates. Finally, in terms of variance, it can be observed in Figure 6 that

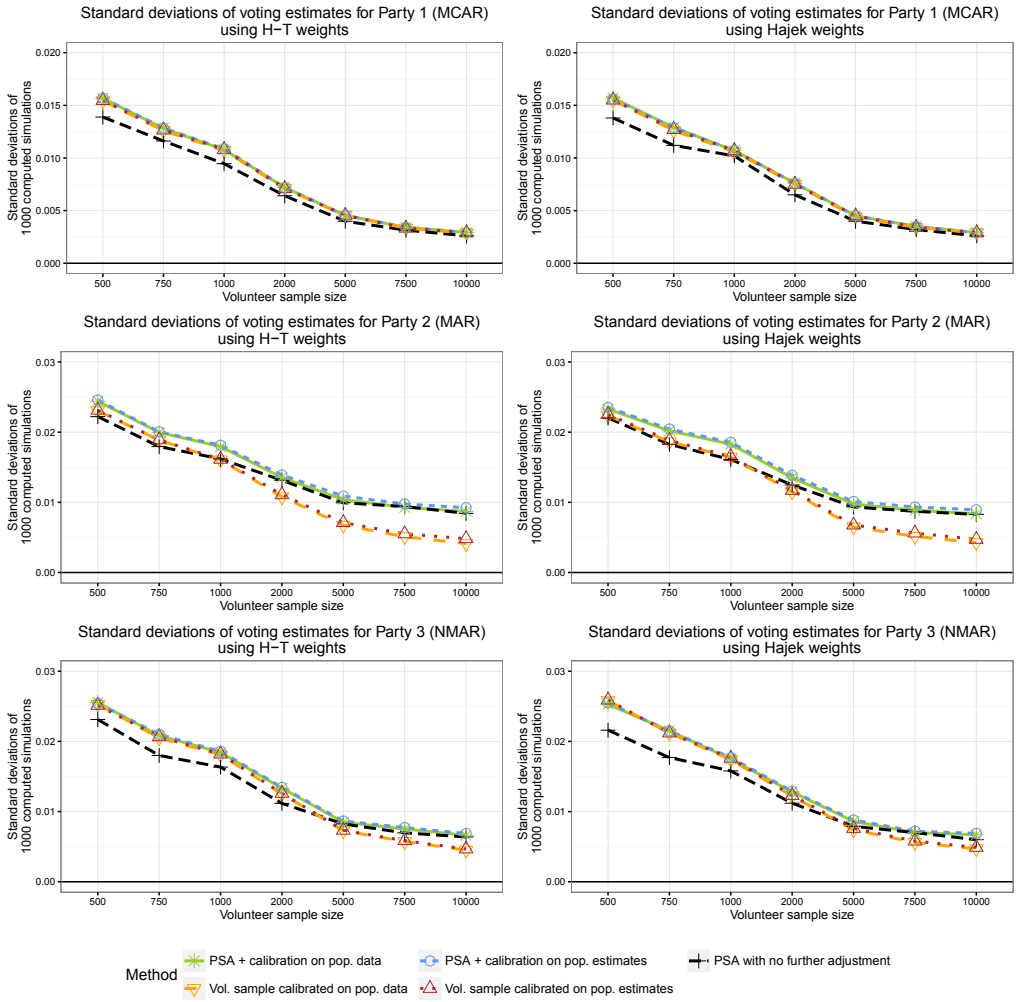


Figure 4: *Standard deviation of voting intention estimations by party provided by each method in Situation 2.*

Hajek-type estimators have a greater variance than Horvitz-Thompson-type estimators, specially when the volunteer sample size is relatively small.

Figures 7 and 8 summarize results for Situation 4. The differences between weighting methods disappear in the MCAR case but remain in the MAR and NMAR cases. In addition, no reduction in bias could be attributed to the calibration of the sample, in contrast with Situation 3, where calibration resulted in less biased estimates in all cases. Regarding standard deviations, the most remarkable result in this situation is the increase in variance that calibration produces in this situation.

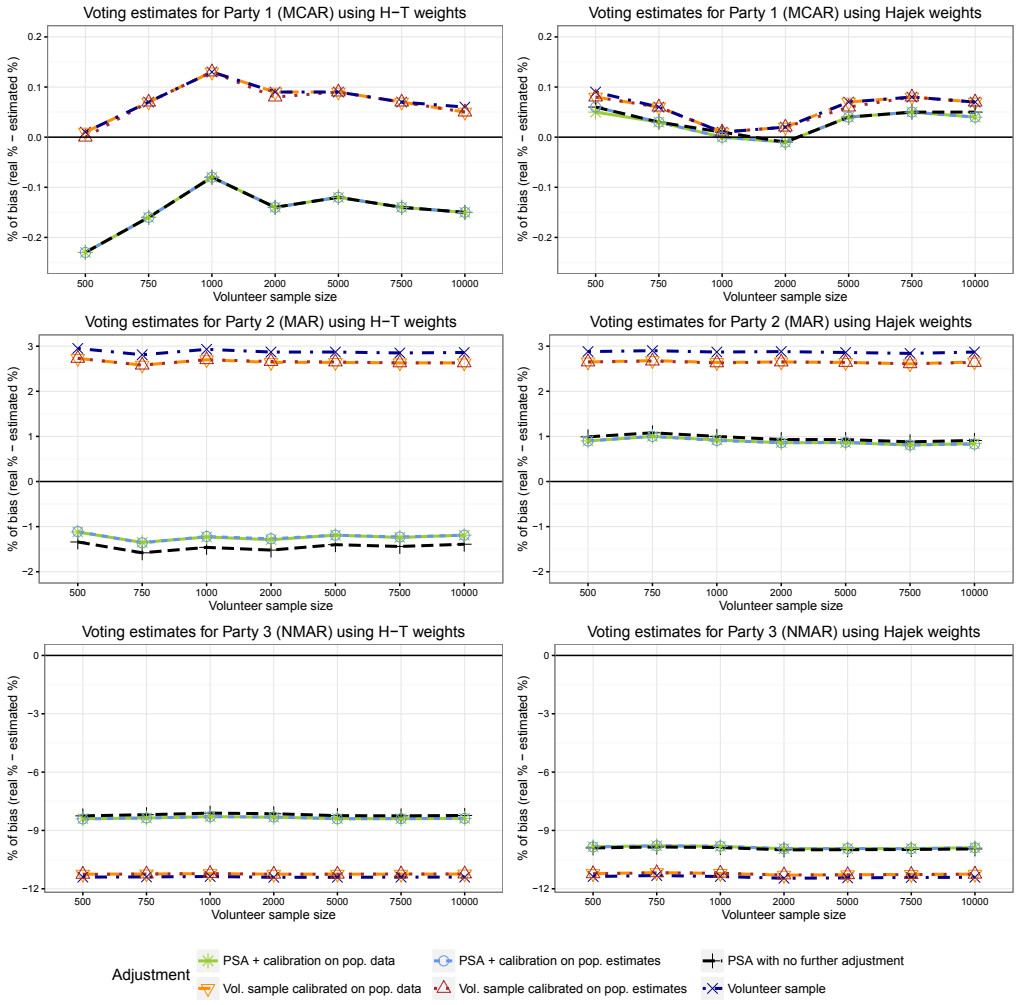


Figure 5: *Bias of each method in voting intention estimations by party in Situation 3.*

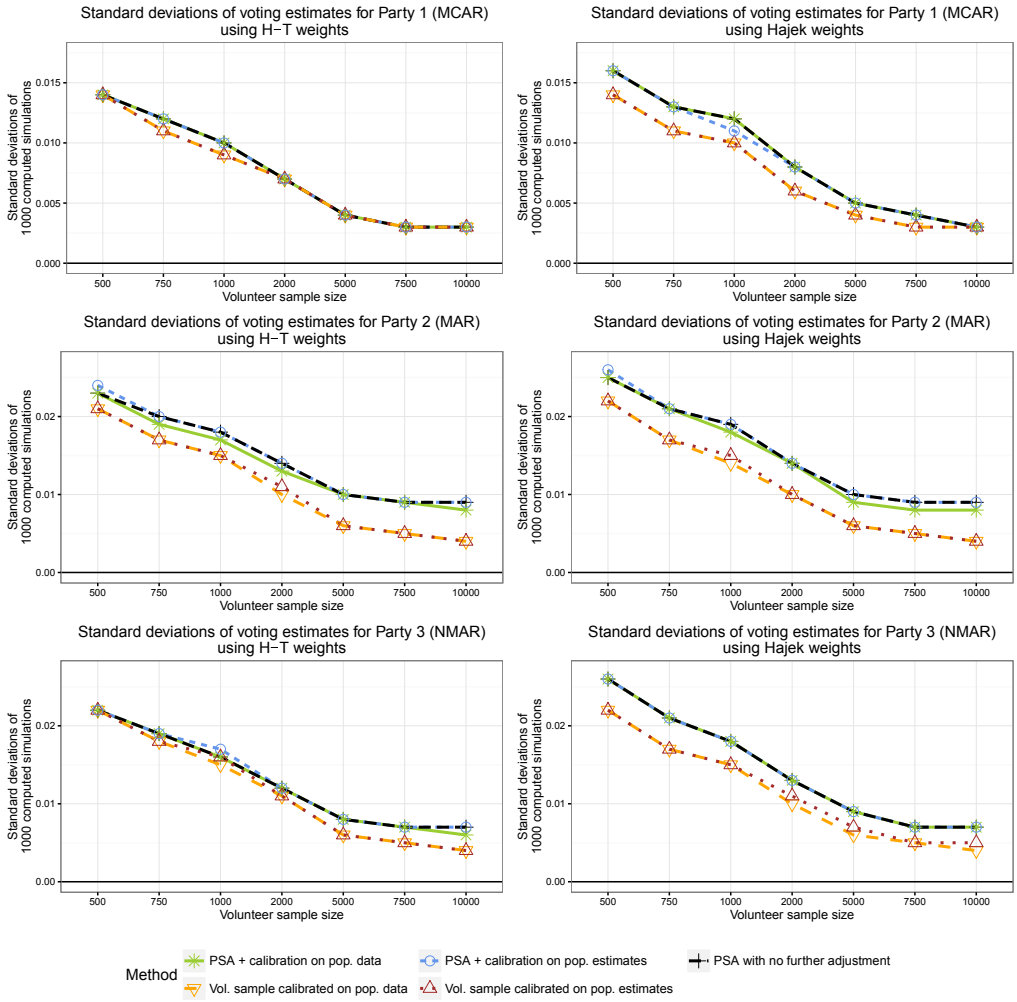


Figure 6: *Standard deviation of voting intention estimations by party provided by each method in Situation 3.*

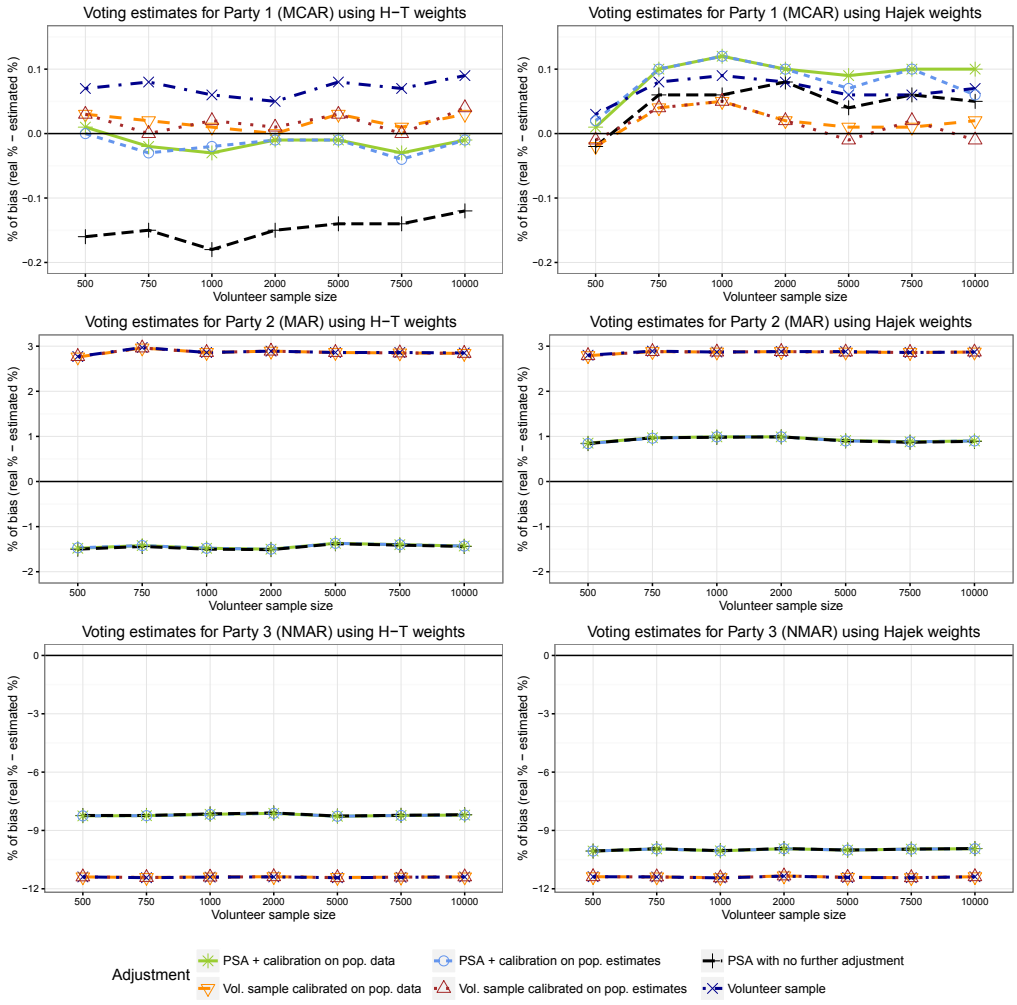


Figure 7: *Bias of each method in voting intention estimations by party in Situation 4.*

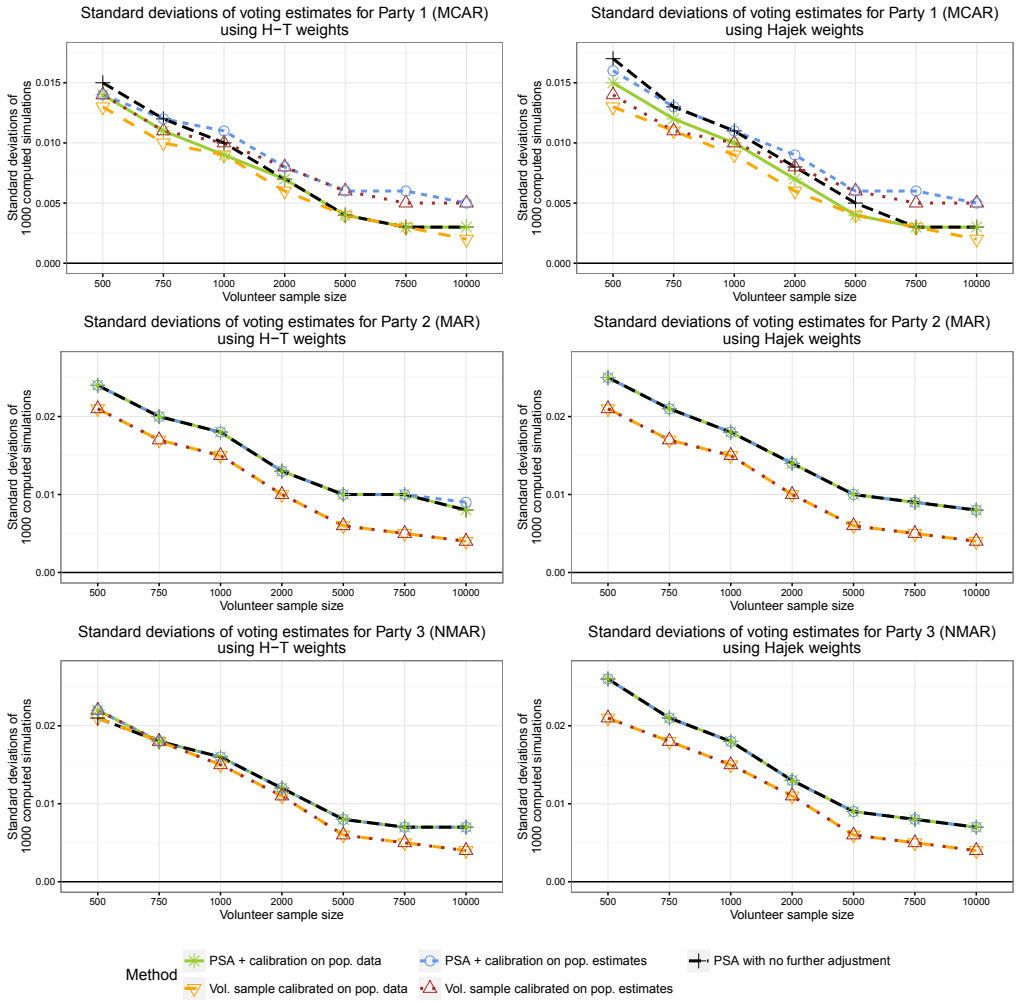


Figure 8: *Standard deviation of voting intention estimations by party provided by each method in Situation 4.*

4. Application study

4.1. Data description

The probabilistic sample data for the application case was obtained through a survey conducted amongst the students of the University of Granada, Spain (UGR) in 2015, with a sample size of $n = 856$ participants. Respondents were recruited through face-to-face interviews following a cluster sampling scheme in three phases, in which Faculties were the primary units, degrees were the secondary units, and academic years were the tertiary units. A total of 34 clusters were randomly drawn from the population

following this design. Sampling error was estimated at $\pm$ 3.3% given the sample size and a confidence level of 95%. Respondents had to complete questionnaires which included several screening instruments for certain kinds of abuse or dependency, including the Cannabis Abuse Screening Test (CAST) and the Severity of Dependence Scale (SDS), which were both validated for the sample. The questionnaire also measured the age and gender of the participants.

The non-probabilistic sample used in this application case came from a survey performed in 2017 by students of the UGR amongst their peers, with a sample size of $n =$ 341 participants. Respondents were recruited following a snowball sampling scheme in online social networks, and completed the questionnaire using an online platform (Google Drive^{TM}). The questionnaire included the CAST and the SDS, as well as questions regarding the age and gender of the respondents. The sampling method implied an internet connection from the respondent and a certain willingness to volunteer in the survey, meaning selection bias came from the same sources than in most of the online non-probabilistic surveys.

The aim of the application was to estimate the SDS mean score for the non-probabilistic sample using the aforementioned correction techniques. Given that SDS scores were provided only for cannabis users in both samples, the original sample sizes dropped out to $n = 115$ participants for the probabilistic survey and $n = 87$ for the non-probabilistic survey.

4.2. Results

The probabilistic sample was used to estimate the total number of cannabis users in the UGR by age groups and gender. These estimates were used as population totals in calibration, in reference to the simulation study results which showed no difference, in terms of bias reduction, between using actual population totals or their estimates. However, this meant that only age and gender could be used as calibration variables. On the other hand, PSA could be performed using age, gender and CAST scores. Differences in data for the three variables between both samples can be consulted in Table 3.

The difference in gender proportions between both samples is statistically significant ($p = 0.0012$), hence it can be assumed that the frames from which samples were withdrawn had different gender proportions. However, this assumption cannot be made for any of the other variables; no practical or statistical significance was found in the difference between samples. These results are an evidence of the lack of discriminant power of PSA potential covariates, thus the propensity of belonging to any of both samples might be much less explanatory.

Estimates of the SDS mean score were computed for each possible combination of techniques (no adjustment, calibration, PSA, and PSA with calibration), auxiliary variables and PSA covariates. Hajek estimator weights were computed in PSA considering the small number of covariates to be used in several combinations, which might not allow to properly allocate the propensity in groups. In each case, jackknife leave-one-out

Table 3: *Means and relative frequencies of each sociodemographic level in the studied samples, and p-values for tests of independence or difference in means performed on each variable.*

Variable	Level	Probab. sample	Non-probab. sample	p-value
Gender				
	Male	51.30 %	74.71 %	0.001[a]
	Female	48.70 %	25.29 %	
Age				
	18 or younger	13.91 %	16.09 %	0.425[b]
	19	13.91 %	18.39 %	
	20	9.57 %	12.64 %	
	21	20.87 %	10.34 %	
	22	12.17 %	14.94 %	
	23 or older	29.57 %	27.59 %	
CAST score				
	Mean score	4.435	5.322	0.167[c]

[a]Two sample test for equality of proportions with continuity correction

[b]Pearson's chi-squared test

[c]Welch two-sample t-test

was performed in order to compute an unbiased estimate of the standard error committed by each method. Results are presented in Table 4, along with the relative difference (in percentage) between each estimate and the mean SDS score provided by the probabilistic sample.

In this application, reweighting with PSA and a Hajek-type estimator is the less biased alternative when using gender, age and CAST score as PSA covariates. When using only gender and CAST scores, the estimator achieves the minimum standard error within all the alternatives. Overall, estimates reweighted with PSA or PSA and calibration to gender and age presented the best results, both in terms of least difference with the reference sample value and least standard error according to the jackknife method.

Table 4: *Estimated SDS mean, standard error and difference with the mean estimated with the probabilistic sample by method, calibration auxiliary variables, and PSA covariates.*

Method	Calibration aux. variables	PSA covariates	Mean SDS score Estimated	Std. Err.	Dif.
Reference sample					
Unweighted			6.261	0.199	
Volunteer sample					
Unweighted			7.264	0.272	16.03 %
Calibration					
	Sex		7.004	0.253	11.87 %
	Age		7.206	0.276	15.09 %
	Sex and age		6.904	0.253	10.26 %
PSA (Hajek)					
		Sex	6.939	0.252	10.84 %
		Age	7.349	0.286	17.39 %
		CAST	6.986	0.246	11.58 %
		Sex, age	6.997	0.266	11.76 %
		Sex, CAST	6.790	0.238	8.46 %
		Age, CAST	6.971	0.251	11.34 %
		Sex, age, CAST	6.742	0.247	7.68 %
PSA (Hajek) + calibration					
	Sex	Sex	7.311	0.278	16.77 %
		Age	7.007	0.253	11.92 %
		CAST	7.028	0.253	12.25 %
		Sex, age	7.323	0.280	16.97 %
		Sex, CAST	7.311	0.278	16.78 %
		Age, CAST	7.052	0.254	12.63 %
		Sex, age, CAST	7.331	0.281	17.10 %
	Age	Sex	7.182	0.283	14.70 %
		Age	7.126	0.264	13.82 %
		CAST	7.239	0.278	15.62 %
		Sex, age	7.086	0.270	13.19 %
		Sex, CAST	7.195	0.282	14.92 %
		Age, CAST	7.136	0.261	13.97 %
		Sex, age, CAST	7.086	0.266	13.18 %
	Sex and age	Sex	7.216	0.283	15.26 %
		Age	6.837	0.243	9.20 %
		CAST	6.955	0.254	11.09 %
		Sex, age	7.136	0.272	13.97 %
		Sex, CAST	7.233	0.283	15.53 %
		Age, CAST	6.875	0.240	9.81 %
		Sex, age, CAST	7.145	0.269	14.12 %

5. Discussion and conclusions

In the last years we are witnessing a strong development of online research methods in general and web surveys specifically. Web surveys are a very attractive option because fieldwork costs are rather low when compared with other modes as mail, telephone and face to face. In addition to cost-effectiveness, there are other reasons that explain why the market research industry has decidedly embraced web surveys in the last years such as the speed of data collection and the advantages associated with the computerization of the questionnaire and self-administration. However, currently the web survey mode has some limitations to adequately represent the general population. In spite of the fast adoption of the internet in the last decades, the number of non-users is still important in most countries. Moreover, non-internet users differ significantly from those who have access and use this technology. As a result, web surveys that fail to include non-internet users are at a high risk of incurring in coverage bias. A second problem that hinders the use of probability sampling in web surveys of the general population is the lack of a proper sampling frame.

In this paper we have focused on the problem of the the lack of coverage of non-probabilistic samples. It is obvious that such a problem can be responsible for a large increase in the bias of the final results. Various correction techniques, such as calibration and Propensity Score Adjustment or PSA, can be applied to remove the bias. This study attempts to analyse the efficiency of correction techniques in multiple situations, applying a combination of PSA and calibration.

The simulation study, which is a technique widely used when studying methods to improve the estimates provided by problematic surveys and particularly calibration or PSA (Lee, 2006, Lee and Valliant, 2009, Kim and Park, 2009, Bethlehem, 2010), is performed in this work with several limitations, such as the variables selected for PSA and calibration and the diversity among possible situations.

Some of the results presented in this work successfully reproduce relevant findings of the existing literature. For example, it is proved in Bethlehem (2010) that bias can be highly reduced through calibration with the right covariates when the non-response due to volunteering has a MAR scheme, while it cannot be equally done in NMAR situations. This is similar to the results obtained in the simulation study; PSA achieves an improvement in the amount of bias much higher for MAR than for NMAR, but as a difference, the right covariates were used for PSA this time rather than for calibration. As a result, calibration fails to remove any bias if not combined with PSA. These results can be linked to Lee (2006), where it was stated that it is critical to add covariates related to the objective of the study, in order to make PSA useful. These findings are relevant in the sense of finding a procedure to remove coverage error when calibration with covariates is not possible; however, results also show that using estimates of the population totals does not cause any significant difference in final results, therefore the usage of the reference survey to estimate population totals of covariates might be considered for calibration purposes.

In addition, it is worth to note that this work introduces the comparison of the efficiency of Horvitz-Thompson and Hajek weights for PSA, a duality proposed in Schonlau and Couper (2017). Results of this study conclude that a difference in efficiency can be made between both approaches only if the right covariates and calibration totals have been chosen previously, and in fact the individual observed differences in weights computed in the simulation study are negligible. This could be explained by the fact that the strata formed with the propensity scores are thought to have individuals whose propensity score is very similar between them, something feasible given the features of the logistic regression model used for that purpose. Under these circumstances, it is very likely that stratification makes no effect in the computation of final weights. On top of that, PSA weights were subsequently used as original calibration weights, contributing to dilute even more the difference between the former.

Finally, the application of the developed adjustment methods in a specific volunteer survey reflects the conclusions of several studies performed in the past on PSA (Lee, 2006, Valliant and Dever, 2011) that the choice of covariates used for the PSA plays a fundamental role on its further efficiency. However, as it happens in most of health-related surveys, this application is limited by the fact that there are no population totals that estimates can be compared with. Further studies should take into account the availability of population counts in their earlier research steps.

On the other hand web surveys, as any other survey, suffer from non-response even if the use of responsive or adaptive design features account for participation rates. Non-sampling errors are particularly important when the investigator has to gather information concerning highly personal, sensitive, stigmatizing and perhaps incriminating issues such as abortion, drug addiction, HIV/AIDS infection status, duration of suffering from a disease, sexual behaviour... In these situations, collecting data by means of survey modes based on direct questioning methods of interview is likely to encounter two serious problems: (i) participants in the survey may deliberately release untruthful or misleading answers, or (ii) participants may refuse to respond ("unit nonresponse" or "item nonresponse") due to the social stigma or because they feel threatened by such inquiries and fear that their personal information may be released to third parties for purposes other than those of the survey.

A considerable limitation of the presented approach could be the "big data" issues that may arise when the volume of data gets larger. This is a feasible situation in internet surveys, given that their characteristics allow for an important number of respondents to take part on them. The main potential limitation of PSA under these circumstances could be related to the adequacy of logistic regression as a predictor for propensity scores, as they would tend to oversimplify the actual relationships between covariates and target variables. The usage of some alternatives to these models, such as machine learning algorithms (e.g., classifiers), should be considered in future research in the area.

Acknowledgements

The authors thank the valuable comments and suggestions given by two anonymous reviewers. This study was partially supported by the Spanish grant MTM 2015-63609-R.

References

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78, 161–188.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615–620.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464–494.

Couper, M. (2017). Developments in survey collection. *Annual Review of Sociology*, 43, 121–145.

Couper, M., Kapteyn, A., Schonlau, M. and Winter, J. (2007). Noncoverage and non-response in an internet survey. *Social Science Research*, 36, 131–148.

Couper, M. and Peterson, G. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, 35, 357–377.

Dever, J. A., Rafferty, A. and Valliant, R. (2008). Internet surveys: can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, 47–62.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87, 376–382.

Díaz de Rada, V. (2012). Ventajas e inconvenientes de la encuesta por internet. *Papers*, 97, 193–223.

Díaz de Rada, V. and Domínguez, J. A. (2015). The quality of responses to grid questions as used in Web questionnaires (compared with paper questionnaires). *International Journal of Social Research Methodology*, 18, 337–348.

Díaz de Rada, V. and Domínguez, J. A. (2016). Mail survey abroad with an alternative web survey. *Quality and Quantity*, 50, 1153–1164.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249–264.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, 21, 111–121.

Kim, J. K. and Park, M. (2009). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21–39.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 329–349.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319–343.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Manfreda, K. L., Berzelak, J., Vehovar, V., Bosnjak, M. and Haas, I. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, 50, 79–104.

Martínez, S., Rueda, M., Arcos, A. and Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233, 2265–2277.

Mei, B. and Brown, G. (2017). Conducting online surveys in China. *Social Science Computer Review*, 0894439317729340.

National Institute of Statistics (2016). Población (españoles/extranjeros) por edad (grupos quinquenales), sexo y año. Retrieved from http://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=02002.px (Accessed 20 March 2018).

National Institute of Statistics (2017a). Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares. Retrieved from http://www.ine.es/prensa/tich2017.pdf (Accessed 20 March 2018).

National Institute of Statistics (2017b). España en Cifras 2017. Retrieved from http://www.ine.es/prodyser/espacifras/2017/index.html (Accessed 20 March 2018).

National Institute of Statistics (2017c). Nivel de formación de la población adulta (de 25 a 64 años). Retrieved from http://www.ine.es/ss/Satellite?c=INESeccionC&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&cid=1259925481659&L=0l (Accessed 20 March 2018).

Pew Research Center (2017). Demographics of Internet and Home Broadband Usage in the United States. Retrieved from http://www.pewinternet.org/fact-sheet/internet-broadband/(Accessed 20 March 2018).

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87–94.

Rueda, M., Sánchez-Borrego, I., Arcos, A. and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71, 33–44.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99–119.

Schonlau, M. and Couper, M. (2017). Options for conducting web surveys. *Statistical Science*, 32, 279–292.

Schonlau, M., van Soest, A., Kapteyn, A. and Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, 37, 291–318.

Pasadas-del-Amo, S. (2018). Cell phone-only population and election forecasting in Spain: The 2012 regional election in Andalusia. *Revista Española de Investigaciones Sociológicas (REIS)*, 162, 55–72.

Taylor, H. (2000). Does internet research work? *International Journal of Market Research*, 42, 51–63.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W. and Terhanian, G. (2001). The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 US elections. *International Journal of Market Research*, 43, 127–135.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105–137.

SORT 42 (2) July-December 2018, 183-206 DOI: 10.2436/20.8080.02.74

Field rules and bias in random surveys with quota samples. An assessment of CIS surveys

Jose M. Pavía[1] and Cristina Aybar[2]

Abstract

Surveys applying quota sampling in their final step are widely used in opinion and market research all over the world. This is also the case in Spain, where the surveys carried out by CIS (a public institution for sociological research supported by the government) have become a point of reference. The rules used by CIS to select individuals within quotas, however, could be improved as they lead to biases in age distributions. Analysing more than 545,000 responses collected in the 220 monthly barometers conducted between 1997 and 2016 by CIS, we compare the empirical distributions of the barometers with the expected distributions from the sample design and/or target populations. Among other results, we find, as a consequence of the rules used, significant overrepresentations in the observed proportions of respondents with ages equal to the minimum and maximum of each quota (age and gender group). Furthermore, in line with previous literature, we also note a significant overrepresentation of ages ending in zero. After offering simple solutions to avoid all these biases, we discuss some of their consequences for modelling and inference and about limitations and potentialities of CIS data.

MSC: 62D05

Keywords: Centre for Sociological Research, quota sampling, fieldwork rules, age and gender groups, inter-quota distributions, intra-quota distributions

1. Introduction

The revolution in information and communication technologies is transforming society and changing the world of business. Collecting, transmitting and storing vast amounts of data (both structured and unstructured) is more viable than ever and has made it easier to find out the individual needs and wishes of certain sectors of the population. As instruments for studying the opinions and attitudes of society as a whole, however, they appear less effective (e.g., Burnap et al., 2016; Jungherr et al., 2017; Kalampokis et al., 2017). Challenges, such as the polarization of opinions expressed on the internet, the

[1] UMICCS, GIPEyOP, Av. Tarongers s/n, Dpt. d'Economia Aplicada, Facultat d'Economia, Universitat de València, 46022 València (Spain).

[2] GIPEyOP, Av. Tarongers s/n, Dpt. d'Economia Aplicada, Facultat d'Economia, Universitat de València, 46022 València (Spain).

Received: May 2018

Accepted: November 2018

division of sociopolitical communities, the difficulties that still arise in the automatic processing of natural language and the biases of connectivity and internet use (e.g., Del Vicario et al., 2017; Ebrahimi, Yazdavar and Sheth, 2017; Mellon and Prosser, 2017) together with the enormous limitations of representativeness still apparent in Big Data (Meng, 2016), point towards a future in which random surveys will continue to be a necessary and fundamental tool for governmental and business decision-making.

Conducting an opinion poll using sampling methods is a very complex process with many interconnected issues, which must respond to well-established principles and methodologies (Groves et al., 2009). The design and planning of a probabilistic survey involves bringing together many facets to create a single tool. Specifying objectives, delineating the target population, designing the questionnaire, choosing the sample, specifying the fieldwork rules and offering guidelines on how the interviewers should act are just some of the tasks that the survey designer must ponder before starting any research, trying to anticipate biases and any problems that may arise (Cea D'Ancona, 2004). The biases from which a survey may suffer are not insignificant, could have multiple origins and can even vary depending on the context (Pavía, Badal, and García-Cárceles, 2016), so they should be avoided wherever possible.

In Spain, surveys carried out by the Centre for Sociological Research (CIS, from its acronym in Spanish: Centro de Investigaciones Sociológicas) and, in particular, its monthly barometers, are a benchmark in the sector of opinion and market studies, thanks to, among other things, the professionalization of its network of interviewers, the size of its samples and the spatial distribution procedures it implements (Pavía and García-Cárceles, 2012). However, there is still room to improve CIS surveys; without cost. In this paper, we focus on analysing the impact of the rules that CIS uses for choosing respondents within the home. In particular, we study the effects that fieldwork rules R1 and R2 have in terms of age distribution within each quota (which we will call intra-quota distributions): (R1) "when in a dwelling there is more than one person who meets the conditions demanded by the quota, the youngest one will be interviewed", and (R2) "if it is impossible to obtain a certain age quota, it can be replaced by one of the age adjacent quotas" (Díaz de Rada, 2005, 2014).

Although the controversy following the 1948 US Presidential Election definitely stated the superiority of probability sampling over (old versions of) quota sampling (Mosteller et al., 1949), market researchers and political pollsters all over the world have continued using quota sampling alongside probability-based methods due to its cost-effective relationship (Vehovar, 1999; Vavreck and Rivers, 2008). Indeed, despite its detractors (e.g., Smith, 1983; Sudman, 1976; Marsh and Scarbrough, 1990), quota sampling is currently the dominant method in online studies through the use of panels (Kennedy et al., 2016) and is gaining even more popularity in traditional telephone and face-to-face surveys due to the growing increase of nonresponse rates and the extra costs they entail (Yang and Banamah, 2014). Certainly, although criticism from the statistical academic community against any practice of quota sampling can be sometimes fierce, the reality is that complex, modern versions of quota sampling, like the one used by

CIS, are closer to probability sampling than to (traditional) quota sampling. The quota sampling design operated by CIS in door to door surveys, often referred to as quota-random sampling or quasi-random sampling (Trochim and Donnelly, 2006), is clearly more refined than the standard quota sampling employed in market research and has been positively used by the academia (Stephenson, 1978).[1]

As discussed in the second section, CIS uses quotas for age and gender to choose interviewees within the household. The number of people to be interviewed from each age and gender group (which we will call inter-quota distributions) is determined exogenously, by design, looking to adjust to a reference (or design) population. The number of people who are interviewed for each age within each quota (intra-quota distributions) is, however, endogenous, depends on chance although influenced by R1 (and R2). Our hypothesis is that within each quota the distributions by age will not be adjusted to the distributions of the target population [2] (from which the respondents are chosen) and that there will be an overrepresentation of the younger subjects as a consequence of R1. Regarding the effect of R2 in the intra-quota distributions, we do not have any hypothesis a priori although, as we will see, its effects are visible. In light of the results, we venture a hypothesis (not evaluable with our data) on how the interviewers interpret R2. Likewise, although our research focuses on the study of intra-quota distributions, we also analyse for completeness whether the empirical inter-quota distributions conform to those designed, reflecting on the possible effect on them of R2.

The analysis was carried out based on the study of the variables of sex, age and province collected from the more than 545,000 interviews completed in the 220 barometers conducted by CIS between January 1997 and December 2016. All the microdata used in this research have been obtained from the CIS Data Bank (`www.cis.es`). The supplementary material lists the numbers of the surveys analysed. The data for the comparisons have been obtained from the Spanish National Institute of Statistics, INE (`www.ine.es`). In the case of intra-quota analysis (the main objective of this research), the comparisons are made on the target populations associated with the collection dates of each barometer and in the case of inter-quota analyses on the reference populations used in each barometer to define quotas. Regarding the latter issue, it should be noted that although the CIS now uses the figures of Spanish residents to define the quotas, until 2015 it used the figures of total residents, with a non-regular update schedule (see Table 1).

The rest of the article is structured as follows. The second section briefly describes the sample design used by CIS in its monthly barometers, paying special attention to the selection of individuals within the household. The third section is devoted to delimiting comparison populations. It specifies the reference and target populations of

1. For example, a version of random-quota sampling, similar to current CIS designs, was used during the seventies (due to budgets constraints) in some waves of the prestigious General Social Survey conducted by NORC.

2. In each barometer, the target population is made up of residents over 18 years of age with Spanish nationality at the time of the survey. In the CIS barometers, the reference and target populations have not historically coincided.

each barometer, details how these have been calculated from the official statistics and outlines our reflections on the consequences of applying R1 and R2. The fourth section introduces methodological issues. In the fifth section, the main results obtained are presented. These are extended and complemented substantially in the supplementary material that accompanies this article. Finally, the sixth section summarises the results, proposes solutions and discusses limitations and strengths of CIS data.

2. The barometers of the CIS. Selection of final units

A CIS barometer consists of a personal survey, conducted monthly in homes (except in August), using a standardized questionnaire whose universe (target population) is the population of 18 years or older resident in Spain and with Spanish nationality. In general, barometers had a designed size of 2,500 individuals[3] with, *theoretically*, all units having the same probability of being selected.

The selection procedure for interviewees is carried out within households, after choosing them through a complex sampling procedure of several stages in which stratified sampling, cluster sampling, random routes and age and gender quotas are applied. First, strata are formed crossing the 17 administrative regions of Spain (CC.AA) with seven categories of habitat size (less than or equal to 2,000 inhabitants, from 2,001 to 10,000, from 10,001 to 50,000, from 50,001 to 100,000, from 100,001 to 400,000, from 400,001 to 1,000,000, and more than 1,000,000 inhabitants)[4]. Once the sample size is spread among the strata (usually by proportional allocation), the next step is to determine how many and which census sections to visit in each stratum. For this, municipalities are chosen at random and, within the municipalities, census sections. Municipalities and sections are selected with probabilities proportional to their different sizes. More details on this procedure can be found in Díaz de Rada (2005, 2008, 2014), Rodríguez Osuna (1991, 2005) and Pavía and García-Cárceles (2012).

Once the interviewers are in place, in the selected census sections, the last stage begins: that of choosing respondents. The interviewees are chosen directly by the interviewers, in the field, through a combined application of random routes and quotas for age and gender. Given the impossibility of having a detailed list of the residents in each household, the interviewers choose interviewees using a series of rules, set in advance, that aim to favour randomness and representativeness. First, households are selected and, then, subjects within the household.

For the selection of households, the interviewer receives a list with the streets (and door numbers) that make up the census section and information about the starting point

3. Since September 2018 there are 3000 individuals.

4. Among other deviations from this norm, it should be noted the positive inclusion of Ceuta and Melilla in the barometers in July 2013 or that, in the definition of the strata, there are certain peculiarities related to metropolitan and insular areas.

of her route. From here, the interviewers visit homes following very detailed criteria (which have suffered some variations over time) related to the routes they must follow, the number of dwellings to visit at each door number, the distances between dwellings and door numbers when an interview takes place, the criteria of election and substitution, etc. (Díaz de Rada, 2005, 2008, 2015). When a dwelling is selected (and its members are willing to collaborate), the last stage begins: that of choosing the person to interview. Only one person is interviewed per household.

For the selection of people within a dwelling the most commonly used methods are either chance (e.g., Kish, 1995; Lind, Link and Oldendick, 2000; O'Rourke and Blair, 1983) or quotas. In the barometers, CIS uses the quota method. Before beginning the fieldwork, the interviewers receive instructions on the number of people of each age and gender (quota) that they must interview in each census section and they only select homes with people who meet the characteristics set in the quotas, until they have completed the survey. This approach means that, as field work progresses, more and more households are discarded as a result of not pursuing quotas that have already been filled (Díaz de Rada, 2014). The CIS considers 12 possible groups crossing gender (Female, Male) with age, classified into 6 categories: 18-24; 25-34; 35-44; 45-54; 55-64; and 65 or more years.

When faced with only one person who meets the quota in the home selected, this person is interviewed. The difficulty comes when there is more than one suitable person representing the quotas still to be filled. In this case, the R1 fieldwork rule of CIS recommends interviewing the youngest person. Given that on average younger people spend more time outside the home, this rule tries to facilitate the work of the interviewers, allowing them to choose the people that are more difficult to encounter in the home first. This rule makes sense when people belong to different quotas but it does not favour a representative selection within each age bracket; which, in line with our hypothesis, we will see causes the appearance of imbalances in the distributions by age of the samples. Together with the previous rule, in case of impossibility or difficulty in encountering a certain quota, CIS allows the quota to be met by replacing it with one of the adjacent quotas.[5]

3. Comparison populations

Through the sample design, CIS aims to replicate the structure of age and gender of its target population, which is made up of the Spanish population aged 18 years or over and resident in Spain. Unfortunately, the target and design populations have not always coincided. Until 2015, CIS exclusively used the statistics of total residents to select municipalities and sections and to determine the size of the quotas. This (historical)

5. The fieldwork rules of CIS only permit alterations in the age quota, never in the gender, and only allow one substitution in each route (Díaz de Rada, 2015).

divergence between the target and the reference (or design) populations has introduced certain deviations in the variables being considered that are worth quantifying[6]. In our research, we focus on another issue: studying the impact of R1 and R2 on the distributions by age collected. Thus, the demographic structures derived from the barometers (empirical inter-quota and intra-quota distributions) are compared with the corresponding population structures associated with each barometer (theoretical distributions of design and target populations). The theoretical distributions are obtained from the official statistics of the Spanish National Institute of Statistics (INE).

On the one hand, for the inter-quota analyses, the population figures used by CIS are employed to determine the size of the quotas in the design of each barometer (reference populations). On the other hand, for the intra-quota comparisons, the population figures of resident Spaniards corresponding to the month of completion of each barometer (target populations) are considered, conditional on the sizes of the design quotas in each survey.

3.1. Inter-quota distributions[7]

Historically, in the 20 years of barometers analysed, we can identify three stages with respect to the demographic structures used for the design of the surveys. The first stage covers the period from 1997 to 2005 (except September 2005) and encompasses 98 barometers, in which CIS used the resident population and the 1995 Municipal Register for deciding the distribution of the sample size, the selection of units (census sections) and the determination of the quotas. In this first stage, the size assigned to each cross of age and gender (quota) in each autonomous community was obtained by multiplying the sample size that corresponded to the autonomous community by its corresponding population structure.

The second stage covered the period from January 2006 to June 2015 (including September 2005). In this period, which included 106 barometers, the census is updated annually (with a schedule not completely regular) taking as a reference the resident population of the latest Municipal Register available from INE: 10 different Registers were used (see Table 1). In this stage, the quota size that corresponds to each autonomous community is determined as in the previous stage, although an adjustment is introduced regarding its distribution among census sections. Previously, they were distributed using a random procedure with restrictions. In this period, they are distributed taking into account the structure by age and gender of the census sections to be visited such that when a particular section has more residents of a certain quota, more people in that census section of that quota will be assigned to be interviewed. This new

6. In Pavía and García-Cárceles (2012) a first approximation can be found in the case of intention to vote in pre-election survey.

7. We appreciate the invaluable help of Valentín C. Martínez (technical advisor of the CIS research department) in answering our questions for the preparation of this section. Any inaccuracy that exists is the sole responsibility of the authors.

Table 1: *Municipal Registers (and reference variables) used in the analysed barometers.*

Barometers (dates)	Register	Variable	Number of barometers
1997.01-2005.12 (except 2005.09)	1995	Total residents	98
2006.01-2006.09 (including 2005.09)	2004	Total residents	9
2006.10-2007.09	2005	Total residents	11
2007.10-2008.10	2006	Total residents	12
2008.11-2009.10	2007	Total residents	11
2008.11-2009.10	2007	Total residents	11
2009.11-2010.09	2008	Total residents	10
2010.10-2011.12	2009	Total residents	14
2012.01-2012.10	2010	Total residents	9
2012.11-2013.06	2011	Total residents	8
2013.07-2014.03	2012	Total residents	8
2014.04-2015.06	2013	Total residents	14
2015.07-2016.05	2014	Spanish residents	10
2016.06-2016.12	2015	Spanish residents	6

Source: Compiled by the authors from personal communication with V. Martínez.

strategy seeks to reduce field work time and to compensate for the increasing rates of non-response (Pavía and Larraz, 2012; Díaz de Rada, 2014).

The last stage began in July 2015 and continues today (up to the time of writing this paper). In this stage, new improvements are introduced regarding the construction of the strata and relative to the reference populations used for the design of the survey. The Register is updated annually (see Table 1), but now CIS uses the total number of residents as reference population to define the strata but the total Spanish population to distribute the sample among the strata. Likewise, the Spanish resident population aged 18 or over is used for the allocation of the quota sizes, not taking into account the stratification for its distribution but only the sections that are going to be visited. That is, quotas are now only representative at the national level.

Although the main objective of this paper is to evaluate the bias introduced by the rules R1 and R2 in the intra-quota distributions (since the adjustment to the inter-quotas distributions should be fulfilled, at least approximately, by design), for the sake of completeness, we have also made comparisons (between theoretical and empirical distributions) for inter-quota distributions in the first part of the results section 5.1. Inter-quota distributions are, in turn, necessary to make intra-quota comparisons when analysing combined age groups (see Figure 3).

Before we can use the theoretical inter-quota distributions employed in the design of each barometer, however, we have to address two questions. On the one hand, the problem posed by the unavailability of the data corresponding to the 1995 Municipal Register, used in the barometers from 1997 to 2005. On the other hand, the possible impact that the so-called rounding effect[8] may have on inter-quota distributions, which

8. The rounding effect emerges when one tries to replicate a population structure of several million people through a sample of, at most, a handful of thousands of people, and manifests itself in the differences that exist between the percentage distributions of the design and sample populations.

may cause significant differences between the inter-quota distributions of the reference populations and the inter-quota distributions actually used.

Both problems are addressed in the supplementary material. The first of the questions is solved by making a comparison between the official statistics of the Municipal Registers and Population Now-Cast estimates of INE, from which we obtain an estimate of residents corresponding to the 1995 Municipal Registers[9]. Regarding the second question, the effect of rounding, we find that, although this is significant at CC.AA level (which would affect the autonomous region inter-quota distributions in the barometers prior to July 2015), it does not appear at the aggregate level, nationally[10]. Given that our comparisons between theoretical inter-quota distributions and empirical quotas are limited to the national ambit, we infer from the analysis that we can use as comparative inter-quota distributions the demographic structures that were used to define the quotas in each group of barometers, which can be calculated directly from the official statistics of INE[11].

3.2. Intra-quota distributions

The inter-quota distributions are determined a priori, by design, with the objective of replicating in the sample the same demographic structure (by groups) that exists in the reference population, although, as we have already seen, the reference and target populations have historically shown certain divergences. In the case of intra-quota distributions, however, there is a correspondence with the target population[12]. The interviewers are instructed to select Spanish residents (in households) aged 18 years or older.

With the limits imposed by the inter-quota distribution of the sample (number of people of each group of age/gender that should be interviewed), the interviewers select individuals within each quota among the resident Spanish population. Therefore, without the effect of R1 and R2, the expected intra-quota structures of the surveyed population should mimic those of the target population; a result that will not be expected when the whole sample is considered together. For the whole sample, the expected structure of the surveyed population will not exactly match that of the resident Spanish population (unless the target and reference populations are equal), since each quota (age/gender group) can have a different relative weight in the design (reference) and target populations.

9. See the section "Estimation of the distribution by age and gender (inter-quota) of the 1995 Register" in the supplementary material.

10. The details that support these conclusions are offered in the section "Analysis of the impact of the rounding effect on inter-quota distributions" in the supplementary material.

11. This result is interesting because it makes the analyses that we have carried out more transparent and easily replicable.

12. It should be noted, however, that Ceuta and Melilla are incorporated into the sampling frame in July 2013. This issue has been taken into account when constructing the intra-quota comparison populations.

Once the previous points have been addressed, we are in a position to define the comparison populations for the intra-quota distributions. Specifically, if we denote by $p^t_{e,s}$ the proportion of Spanish residents with age e, sex s at time t, and by $\pi^t_{g,s}$ the proportion of people who, by design, should be selected within the age group g with sex s at time t, we would have that, without the effect of R1 or R2, the proportion (relative size) of the expected intra-quota of Spanish residents with age e, sex s at time t for the population as a whole, $Cp^t_{e,s}$, will be given by equation (1)[13], and that, on the other hand, the expected intra-quota proportion of Spanish residents within the corresponding quota, $Dp^t_{e,s}$, will be given by equation (2).

$$Cp^t_{e,s} = \frac{\sum_{g\in G} p^t_{e,s}\cdot \delta^t_{e,s}\cdot \pi^t_{e,s}}{\sum_{g\in G}\sum_{e^*\in E_g} p^t_{e^*,s}\cdot \delta_{e,g}} \cdot \frac{1}{\sum_{g\in G}\pi^t_{g,s}} \tag{1}$$

$$Dp^t_{e,s} = \frac{p^t_{e,s}}{\sum_{g\in G}\sum_{e^*\in E_g} p^t_{e^*,s}\cdot \delta_{e,g}} \tag{2}$$

where $s \in S = \{W\text{: Women}, M\text{: Men}\}$, $e \in E = \{18,19,20,\ldots,84,85+\}$[14], $g \in G = \{18-24, 25-34, 35-44, 45-54, 55-64, 65+\}$, E_g denotes the set of ages included in the group g (for example, $E_{18-24} = \{18,19,20,21,22,23,24\}$), and $\delta_{e,g} = 1$ if $e \in E_g$ and, otherwise, equal to zero.

The previous notation is complex but its meaning is easy to understand by looking at the following example. Supposing that (i) the percentage of Spanish women residents (target population) of 18, 19, 20, 21, 22, 23 and 24 years old at a given moment are respectively, 0.8%, 0.7%, 1.0%, 1.3%, 1.1%, 1.2% and 1.5% $(p^t_{18,W},\ldots,p^t_{24,W})$ (ii) the total percentage of women resident in Spain between 18 and 24 years old is 7.0% $(\pi^t_{18-24,W})$ and (iii) the total of residents of 18 years or more in Spain is 85% $(\sum_{g\in G}\pi^t_{g,s})$, then it can be seen that, without the effect of R1 and R2, the expected proportion of women aged 20 in the total population interviewed, $Cp_{20,W}$, and the expected proportion of women aged 20 within the corresponding quota in the surveyed population, $Dp_{20,W}$, would be:

$$Cp_{20,W} = \frac{1.0\%\cdot 7.0\%}{0.8\%+0.7\%+1.0\%+1.3\%+1.1\%+1.2\%+1.5\%}\cdot\frac{1}{85\%} \approx 1.08\%$$

$$Dp_{20,W} = \frac{1.0\%}{0.8\%+0.7\%+1.0\%+1.3\%+1.1\%+1.2\%+1.5\%} \approx 13.16\%$$

In other words, these proportions are no more than (i) the relative weight that each age represents in the target population, re-weighted by the design weight of the group

13. Note that $p^t_{e,s}/\sum_{e\geqslant 18} p^t_{e,s} = Cp^t_{e,s}$, if the target and reference populations coincide.

14. In the intra-quota comparisons, people of 85 years or more (85+) have been added to simplify the graphical presentations and to facilitate the verification of the theoretical conditions that are required by the hypothesis tests of goodness-of-fit implemented.

to which it belongs, and (ii) the relative weight that each age represents within its group (quota). Obviously, when working with the subset of men or women, the corresponding Cp proportions would be calculated in a similar way.

In general, the comparisons are made within each quota (Dp proportions), as the results are sharper and easier to interpret. For an overview — i.e., when we compare the empirical and theoretical (expected) distributions of all ages together — we will use Cp proportions, since the distributions equivalent to those collected in the barometers are obtained by re-weighting the intra-quota distributions of the target populations with the inter-quota distributions actually used in the design of each barometer.

In fact, the demographic structure defined by the set of proportions $\{p^t_{e,s}/\sum_{e\geqslant 18} p^t_{e,s}\}$ should not be used as a population comparison to evaluate the impact of rules R1 and R2 on the set of empirical intra-quota proportions since part of the discrepancies that could be observed would not be attributable to the effect of R1 and R2 but would be a consequence of the possible differences existing between $\pi^t_{g,s}$ and $\sum_{e^*\in E_g} p^t_{e,s}\cdot\delta_{e,g}$. A different issue would be to study to what extent the distribution by age and gender of the responses collected in the barometers matches the corresponding structure of the target population, regardless of whether the discrepancies originate from the use of R1 and R2 rules or from the differences between target and design populations. In that case, it would be pertinent to use the set of proportions $\{p^t_{e,s}\}$, duly weighted by restricting to persons of 18 years of age or older, $\{\sum_s p^t_{e,s}/\sum_{e^*\geqslant 18} p^t_{e^*,s}\}$, or men or women of 18 years of age or older, $\{p^t_{e,s}/\sum_{e^*\geqslant 18} p^t_{e^*,s}\}$. For completeness, we also carry out this comparison (see bottom panel of Figure 3). In fact, we think that the differences between the discrepancies of both comparisons could be interpreted as an indicator of the practical impact of the divergences between target and design populations.[15]

To calculate the comparable intra-quota distributions, however, we need to know the distributions of the target populations in each of the months in which the barometers were carried out. The difficulty lies in the fact that Spanish resident population statistics are only published once a year (referenced, except in 1996, to January 1), so we have proceeded to estimate them in each of the months where official statistics are not available. Specifically, we estimate the number of resident Spanish people with age e, sex s in a month τ, $N^\tau_{e,s}$, located between two Municipal Registers referenced in the months t and t^*, with $t<\tau<t^*$, as a weighted average of the corresponding official figures, $N^t_{e,s}$ and $N^{t^*}_{e,s}$, with inverse weighting coefficients to the number of months away. The mathematical expression of the estimator is given by equation (3).

$$N^\tau_{e,s}=\frac{N^t_{e,s}\cdot(\tau-t)^{-1}+N^{t^*}_{e,s}\cdot(t^*-\tau)^{-1}}{(\tau-t)^{-1}+(t^*-\tau)^{-1}} \tag{3}$$

15. Obviously it would be simple to make a direct comparison between the intra-quota distributions of the target and design populations, for which no data needs to be collected. However, these discrepancies would only highlight the impact of the differences between these populations and would not take into account the possible interactions with R1 and R2.

The solution to equation (3) is equivalent to assuming a uniform distribution of demographic events between two consecutive censuses (Lledó, Pavía and Morillas, 2017).

3.3. Intra-quota and inter-quota distributions. Discussion

At this point and before making the comparisons, it is time to think over on the different nature of intra-quota discrepancies (understood as the deviations between the theoretical percentages of people that would be expected by chance to be interviewed within each group of age/gender and those that were really interviewed) and inter-quota discrepancies.

The deviations, statistically significant, in the intra-quota distributions will be an indicator of a non-representative selection of subjects within each quota and of an effect of the rule R1 (and perhaps also of R2). On the other hand, inter-quota deviations theoretically should not exist, since the sample design imposes to the sample to mimic the reference population. There are several reasons, however, why there could also be divergences in the inter-quota distributions, that is, between the sizes of the reference and collected quotas. Discounting (at national level) the possible rounding effect, the discrepancies could occur as a consequence of rule R2.[16] If this rule manifests any effect, however, it should not be very pronounced because the random substitutions between adjacent quotas should occur sometimes in one direction and sometimes in the opposite direction. In any case, given that in practice it is easier to find older people in households (Díaz de Rada, 2014), we expect a greater propensity to make substitutions towards higher age quotas and, therefore, a slight underrepresentation of the youngest quotas.

Regarding the empirical intra-quota distributions, we would initially expect that, at random, the number of people interviewed within each age group would replicate the corresponding distribution of the quota, with the relative higher probability of older people being at home not playing any significant role here, given that within each quota the age differences are relatively small (except for the group of 65 and older). However, due to rule R1, which recommends interviewing the youngest person if there is more than one person suitable for the survey in the home, our a priori is that within each quota there will be overrepresentation of younger people.

4. Methodology

Our main hypothesis is that as a consequence of R1 there will be an overrepresentation of younger people within each quota. This implies, on the one hand, that the proportion of people interviewed with the minimum age of each quota (18, 25, 35, 45, 55 and 65

16. The possible effect of non-response is negligible here. In the 220 barometers analysed, 99.42% of the 550,000 planed interviews (546,789) were carried out.

years) will tend to be greater than the corresponding proportion in the target population and, on the other hand, that the distribution of people surveyed within each quota will not adjust to the theoretical distribution of the target population. Regarding the inter-quota distributions, our hypothesis is that these will be adjusted to the designed distributions, that is, to the distributions of the reference populations used to determine the quotas during the planning of the surveys.

To evaluate these hypotheses we have used tests of hypothesis. On the one hand, unilateral parametric hypothesis tests for a proportion, with null hypotheses $p_m^{\upsilon} \leq p_m^t$ and alternative hypotheses $p_m^{\upsilon} > p_m^t$, where p_m^{υ} represents the true probability that a person of minimum age (m = 18, 25, 35, 45, 55, 65) in the corresponding quota is interviewed according to the fieldwork rules and p_m^t the theoretical proportion of people with that same age in the target population. On the other hand, classical goodness-of-fit χ^2 tests, where the null hypotheses postulate that the empirical distributions conform to the theoretical one and the alternative hypotheses state that they do not fit. Given the tendency of hypothesis tests, especially parametric ones, to accept null hypotheses, the rejection of these, mainly of $p_m^e \leq p_m^t$, will provide strong evidence in favour of our hypothesis regarding intra-quota distributions.

It is known, however, that the classic goodness-of-fit tests (χ^2, Kolmogorov-Smirnov, Kuiper, ...) have a lot of statistical power, so they tend to reject the null hypothesis when the sample size is extremely large (Badal-Valero, Alvarez-Jareño and Pavía, 2018). Hence, as an alternative to the classic χ^2 test, we have also implemented, when working with very large samples, the equivalent Monte Carlo test. This test has as a limit, when the sample size tends to infinity, the uniformly most powerful test associated with the hypothesis (Hope, 1968).

In addition to the results of the statistical tests, which we analyse through p-values, we have also used a classic indicator of dissimilarity to numerically evaluate the degree of adjustment between empirical and theoretical distributions. Specifically, we have employed the mean absolute statistical error, Δ, to measure the average percentage difference between distributions. For the inter-quota comparisons the statistical mathematical expression is given by equation (4) and for the intra-quota comparisons by equation (5)[17].

$$\Delta = \frac{1}{12} \sum_{s \in S} \sum_{g \in G} |\pi_{g,s}^{t*} - \widehat{\pi_{g,s}^{t*}}| \tag{4}$$

$$\Delta = \frac{1}{|E_g|} \sum_{e \in E_g} |Dp_{e,s}^t - \widehat{Dp_{e,s}^t}| \tag{5}$$

where $\pi_{g,s}^{t*} = \pi_{g,s}^t / (\sum_{s \in S} \sum_{g' \in G} \pi_{g',s}^t)$ represents the theoretical percentage that, by design, corresponds to the quota of group g and sex s, reweighted so that the sum of all

17. When working with Cp proportions, equation (5) is of course modified.

the quotas is 100%, $\widehat{\pi^{t*}_{g,s}}$ is the corresponding value in the sample, $\widehat{Dp^{t}_{e,s}}$ is the sample estimation of $Dp^{t}_{e,s}$ and $|E_g|$ represents the number of ages included in E_g.[18]

The presentation of results using only statistical summaries is very useful and succinct, but it can also obscure other interesting facts and make communication of the results less agile. Therefore, in line with recommendations made in Cleveland (1993), we decided to make also comparisons from a graphical perspective. This has made it possible to discover aspects present in the data that could otherwise have gone unnoticed. In our opinion, the power of the graphic representations (presented in the article and in the supplementary material) is such that they will convince the most sceptical reader of the conclusions derived from this study. All statistical calculations and analyses have been performed in version 3.3.1 of R (R Core Team, 2016).

5. Results

This section shows the main results of the comparisons made to study the effect of the fieldwork rules R1 and R2 on the age distributions collected in the 220 monthly barometers carried out by the CIS between January 1997 and December 2016. In addition, the supplementary material, which accompanies this article, considerably broadens the analyses presented in this section, especially in relation to intra-quota distributions.

In the first subsection of this section, a comparison is made between the theoretical and empirical inter-quota distributions, which are expected to show no statistically significant differences. In the second subsection, the analysis of intra-quota distributions is discussed; this shows a greater and richer range of results.

5.1. Inter-quota analysis

As can be seen in Figure 1, where the comparison between the empirical and theoretical inter-quota distributions are grouped according to the Municipal Register used in each barometer to determine the quota sizes, both sets of distributions are very similar, as expected. The greatest differences between distributions are detected for the barometers that used the 1995 Municipal Register for the sample design. This is not surprising since the data actually used in the barometers of that stage are not available and, as explained in the third section and detailed in the supplementary material, they had to be estimated from the Population Now-Cast estimates.

In fact, except for the data set associated with the surveys made with reference to the 1995 Register, the null hypothesis of adjustment of the empirical data to the corresponding theoretical distribution is not rejected for any other set of observations with the usual level of significance (α) of 5%. When the responses of all the barometers

18. The number of ages in E_g equals to 7 when $g = 18-24$, equals to 10 when $g \in G = \{25-34, 35-44, 45-54, 55-64\}$ and equals to 21 when $g = 65-85+$.

Figure 1: *Comparison of distributions between theoretical (design, INE) and empirical quotas (collected in the corresponding barometers, CIS). The comparisons have been grouped according to the Register used to determine the quotas for each survey. As a summary, the last two panels offer an aggregate comparison, with and without the barometers that used the 1995 Register as the reference population. As well as the graphical comparison, each panel shows the p-value associated with the χ^2 goodness-of-fit test, the number of observations used (sample size, n) and the value of the dissimilarity statistic Δ defined in equation (4).*

are combined and compared with the population obtained as a weighted average of the different theoretical populations, the null hypothesis is rejected however. This occurs despite the fact that the dissimilarity statistic value Δ for the set of barometers associated with the Registers from 2004 to 2015 is very small, with an average discrepancy of only 0.08%. This rejection may be due to the known propensity that classic goodness-of-fit tests show to reject the null hypothesis when the sample size tends to infinity. Using the equivalent Monte Carlo test we obtained a p-value of 0.0112, which would not result in a rejection of, in this case, the null hypothesis for $\alpha = 1\%$. The same exercise for the set of observations associated with the 1995 Register or the set of all the Registers, however, continues producing p-values smaller than 0.0001 and indicating the need to reject the null hypothesis in these cases.

Although in terms of joint distributions the R2 rule does not show any effect that represents a statistically significant overall deviation of the theoretical inter-quota distributions, a result is visible in the data, which had already been foreseen in subsection 3.3, related to the existence of a certain tendency towards underrepresentation of the lower age quotas with respect to the designed sizes. This result is consistent with the greater difficulty that exists in practice of encountering younger people to interview and the consequent greater likelihood of substitution for the adjacent higher age quota, as suggested in the application of R2. In fact, we find in 27 of the 30 analysed quotas corresponding to groups of 18 to 24 years old (90%) that the percentage of respondents is lower than the percentage designed. In the case of women, the quota is underrepresented, on average, by 0.25%. This figure is reduced to 0.11% in the case of men, increasing to 0.18% when excluding the three cases in which the proportion of respondents belonging to the quota of men between 18 and 24 years is above the proportion designed. This trend also manifests itself in the following quotas of groups between 25 and 34 years, whose percentage of underrepresentation reaches 70%. Overrepresentation is predominant for the rest of the groups.

5.2. Intra-quota analysis

Just as expected, the sizes of the inter-quota distributions collected fit to the sizes designed in the barometers, with the possible deviations introduced by R2 not having statistically significant effects. In the same vein, the results for the intra-quota distributions are also as expected, in this case, for not adjusting to the theoretical distributions. In fact, as shown in Figure 2, where the comparison between empirical and theoretical intra-quota distributions for the set of barometers is shown graphically, the null hypothesis, which states that within each quota the empirical distributions by ages conform to the theoretical ones, is rejected for all age and gender groups.

Added to the generalized rejection of the hypothesis of equality between empirical and theoretical intra-quota distributions is the expected rejection of the hypothesis that, in each quota, the observed proportion of respondents with the minimum age of the quota is equal to or less than the corresponding theoretical proportion. This rejection,

which is to be expected as a logical consequence of the use of the R1 rule, provides evidence to support our hypothesis. Looking more closely at the different representations shown in Figure 2 (and also in Figure 3, where all the information contained in the different panels of Figure 2 is presented together), the previous conclusions are not the only interesting conclusions to be made. In addition to the sample overrepresentation that is apparent for the minimum ages of each quota, two further results stand out. On the one hand, it is clear that for each age/gender group, the upper limits of each quota are also overrepresented. On the other hand, similarly, the ages ending in zero are also generally overrepresented, excluding 20 years[19].

The existence of overrepresentation of the 'whole' ages (finishing in zero or in five) in censuses and surveys has been a recognized dysfunction for many years (e.g., Myers, 1940; Bachi, 1951 and Carrier, 1959), which usually manifests itself more strongly in populations with lower levels of education, as various studies on historical populations or more recent transnational analyses have shown (Brian, Baten and Crayen, 2009; Lyons-Amos and Stones, 2017). In this literature, the numbers ending in zero and, to a lesser extent, in five are seen to act as poles of attraction (rounding) for the interviewees when declaring their age, causing relative gaps in the adjacent ages. These same patterns are also observed, in part, for the ages ending in zero in the empirical age distributions of the CIS barometers, especially after age 40. In our data, a greater gap is observed in the adjacent older ages than in the adjacent younger ones, which could suggest that the rounding responds to a psychological self-deception on the part of the interviewees who wishes to consider themselves younger or to a certain effect of social desirability in an attempt to appear younger before the interviewer.

For ages ending in five, which coincide with the lower limits of our intervals (except for the lower age quota), we also observe overrepresentation, although the same previous patterns as in ages ending in zero are not observed in Figure 2 or, if anything, to a small degree for ages 45 and 65. For these ages, the fact that the ages 46 and 66 show values below 47 and 67, respectively, would suggest that a part of the overrepresentation observed for the minimum ages of these intervals could be a consequence of the attraction effect towards 'whole' ages. Fortunately, for our hypothesis regarding the impact of R1, two results clearly point out that the distortion effect of R1 in the intra-quota distributions is genuine and of appreciably greater intensity. On the one hand, the overrepresentation shown by the age(s) 18 (and 19) cannot be attributed to the effect of attraction towards 'whole' ages. On the other hand, our data show an overrepresentation of ages ending in five consistently greater than the overrepresentation of ages ending in zero (see also Figures S3 to S137 in the supplementary material), whilst the literature states the overrepresentation of ages ending in five as being systematically of less intensity.

19. Although in Figure 2 for the 30 years age group there is no overrepresentation in the strict sense, it is notable that the percentage of people who claim to be 30 years old is noticeably greater than those who claim to be 29 or 31 years, which is a clear indicator of relative overrepresentation.

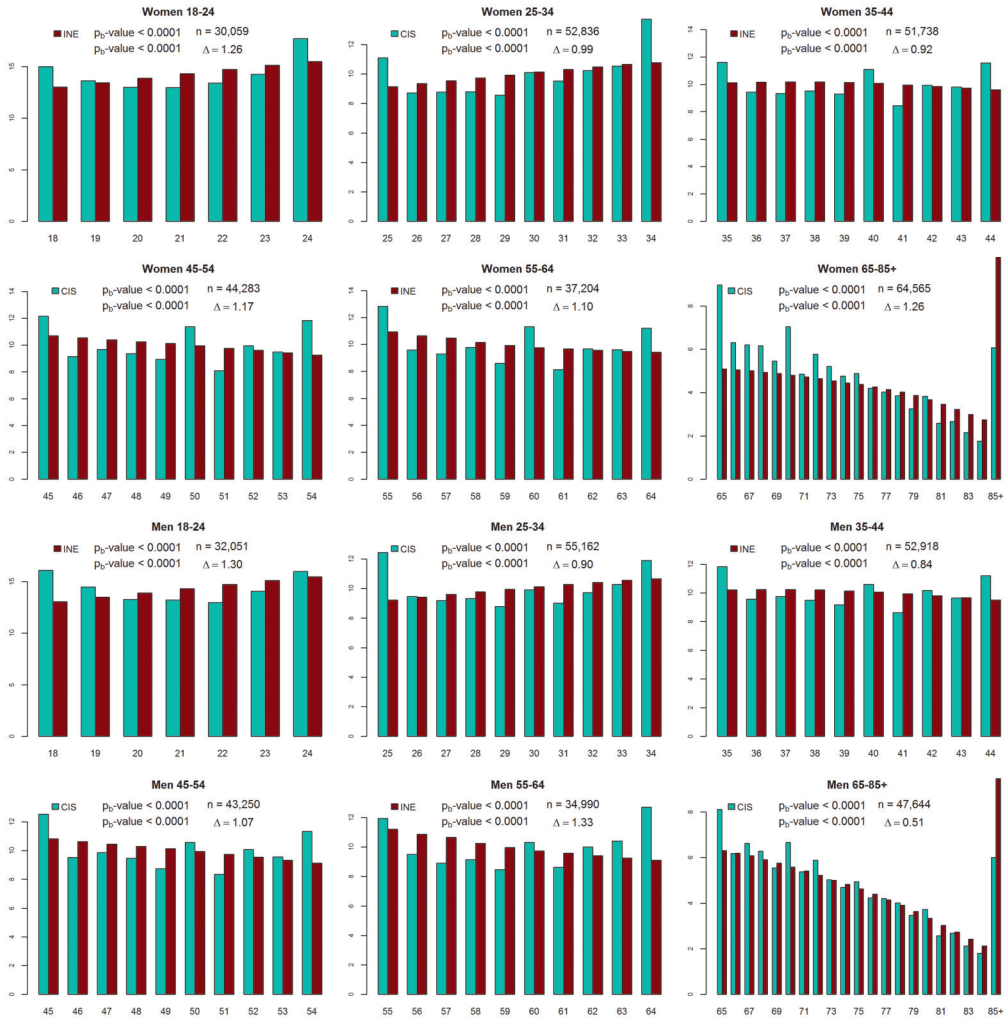


Figure 2: *Comparison between the theoretical intra-quota distributions (in the target population, INE) and the empirical distributions (set of responses collected in the 220 barometers analysed, CIS). The theoretical distributions have been calculated as the sum of the theoretical distributions associated with each barometer. In addition to the graphical comparison, each panel shows the p_b-value associated with the χ^2 goodness-of-fit test, the p_m-value associated with the unilateral test for the minimum proportion of each quota, the number of observations used (size of the sample, n) and the value of the dissimilarity statistic* Δ *defined in equation (5).*

Regarding the other surprising result, that of the huge overrepresentation shown by the upper limits of all quotas of age and gender (except obviously in the groups of 65 and over), our assumption (impossible to contrast with our data) is that this is due to the way the interviewers apply the R2 rule. Our assumption would be that, in order to minimize their work time, the interviewers are more willing to use the R2 rule as the field work progresses. Specifically, we conjecture that when interviewers have fewer and fewer

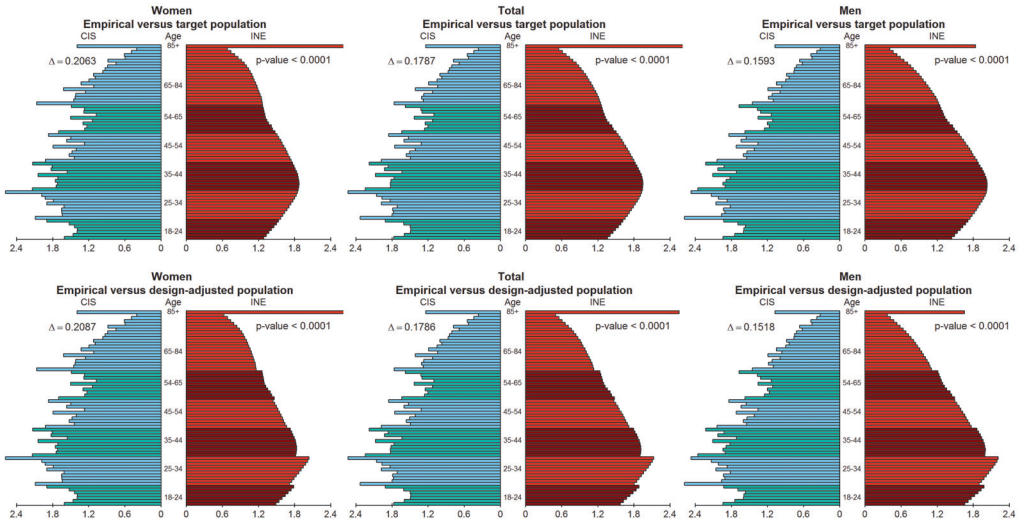


Figure 3: *Comparison between the empirical percentages of interviewees of each age group (for women, men and total) for the set of barometers (CIS) and the theoretical percentages expected according to the target and design population (INE) of the survey (upper panel) and the target population (bottom panel). In the panels corresponding to women and the total, the right axis is truncated (the percentage of people with 85 years and more is higher than shown) in order to make the results clearer. The theoretical percentages are calculated as a weighted sum of the theoretical percentages corresponding to each stage and each barometer. In addition to the graphical comparison, each panel shows the p-value associated with the* χ^2 *goodness-of-fit test and the value of the dissimilarity statistic* Δ *defined by equation (5) adapted to the* Cp *proportions.*

free quotas available they tend to take advantage of the R2 rule to fill a missing quota with a person whose age is one year either side of that quota as soon as s/he is available. This assumption would explain the observed overrepresentation of the upper limits and would help to explain part of the overrepresentation of the lower limits of the different quotas, except, for obvious reasons, in quotas of ages from 18 to 24.

The alternative explanation according to which the overrepresentation of the upper limits within each quota could be mainly due to the positive correlation that exists between age and time spent in the home does not hold. On the one hand, a big difference is seen between ages ending in three and those ending in four and, on the other hand, the figures for ages ending in two and in three are practically the same.

The observed patterns in the data set as a whole do not seem to be a consequence of isolated operations or trends, occurring only at certain moments of time or in a part of the territory, but to a generalized global trend that extends across all provinces, autonomous communities, years and stages, as can be seen in Figures S3 to S137 and in Tables S3 to S6 of the supplementary material. For example, the average overrepresentations at provincial level in the intra-quota percentages of men aged 35, 40 and 44 are, respectively, 1.6%, 0.5% and 1.7%, affecting 80.8%, 59.6% and 76.9% of the provinces, respectively (see Table S4). This result, along with many others visible in the

data (see Figures S3 to S104 and Tables S3 to S6), would reinforce the likelihood of our conjecture about how interviewers apply R2.

Finally, by performing a combined analysis of the intra-quota distributions, we see, as expected, that the empirical distributions are more adjusted to the populations defined by the Cp proportions (that is, incorporating the differences between the design and target inter-quota distributions) than those of target populations. Although this is not entirely evident by comparing the values of the dissimilarity coefficients Δ of the upper and lower panels of Figure 3, this result becomes unquestionable when analysing Figures S105 to S137 of the supplementary material. For example, considering exclusively the values of Δ for the years 2006 to 2016,[20] we see that, on average, the dissimilarity value Δ is 11.8% higher when the combined empirical distributions are compared with the target populations (upper rows of the panels) than when compared with the adjusted-target populations (lower rows of the panels).

6. Summary and concluding remarks

Due to budget and time constraints, quota samples are extensively used by opinion pollsters and consumer researchers all over the world. Quota sampling, however, is strongly contested by the statistical community due to its break from randomness. Hence, to achieve representative samples, the more prestigious organizations relying on this approach try to follow wherever possible random steps and rules in their sampling designs and only use quota sampling in the final step. CIS is one of these organizations. Their surveys are far removed from traditional quota sampling, being closer to probability sampling. Indeed, surveys carried out by CIS, in particular its monthly barometers, are a point of reference for Spanish market research and public opinion polls.

Whatever the approach adopted, however, we should not overlook the detail. This paper shows that even seemingly innocuous fieldwork rules can have significant consequences, biasing the collected samples. Our analysis tries to warn practitioners about the need of thinking carefully about all the components that define a sampling design. We show that, for selecting individuals within the household, the CIS fieldwork rules R1 (of selection, in case of doubt, between and within quotas) and R2 (of substitution, in case of difficulty, between quotas) should be improved, since they lead to biases in the distribution by age within each quota. In this article we study, using the more than 545,000 responses collected in the 220 barometers conducted between 1997 and 2016, the effect of these rules on age distribution and analyse the existence of bias comparing the empirical distributions with the theoretical ones (those expected according to the sample design and/or the target populations).

20. An assessment excluding the barometers designed using the 1995 Register is pertinent as it avoids the possible distorting effect that could introduce in the analysis the use of estimated data for inter-quota distributions associated with the barometers referenced to the 1995 Register.

The study implemented reveal a set of highly interesting results that support the existence of biases in the intra-quota distributions of the CIS barometers. Regarding the inter-quota distributions, it is observed that rule R2 induces a certain tendency for underrepresentation of younger quotas, consistent with the lower probability of encountering younger people in the home, although this causes no statistically significant deviation. The most interesting results are seen for the intra-quota distributions.

Specifically, within all quotas the hypothesis that the empirical distribution by ages fits the theoretical distribution is rejected, illustrated by the three overrepresentations highlighted. Firstly, as a logical consequence of the use of the R1 rule, the results show that there is a significant overrepresentation of the proportion of interviewees in the survey having the minimum age of each quota. Secondly, as has been previously observed in dozens of studies, we detect the existence of a significant overrepresentation of ages ending in zero. Thirdly and unexpectedly, it is notable that the maximum ages of each quota (excluding older quotas) are clearly overrepresented.

Regarding the last overrepresentation, our speculation is that this is a consequence of R2. We surmise that, shielded by rule R2, interviewers tend to choose a person from an adjacent quota when the person to be interviewed is only one year of age away from a quota not covered. If this conjecture were true, this would also help to explain part of the overrepresentation observed in the minimum ages of each quota. Obviously, our previous supposition is questionable. We could ask ourselves why, if R2 allows interviewers to substitute respondents with whatever age of an adjacent quota, do they decide, as a rule, to restrict themselves to just one year either side of the quota. We venture that maybe, just to stop what they can anticipate it will be a hard search, they are reluctant to discard an encounter when they find a willing respondent who is only a year away from the preferred quota, believing that this substitution would not represent a significant deviation. In order to confirm (or reject) our supposition, we would need to know in each census section the order followed to complete all the interviews. This will allow us to build the ordered distribution of ages of the interviewees and to assess the likelihood of our assumption. Furthermore, having access to the "sampling sheet" would also allow us to check if interviewers follow the rule of performing a quota-age substitution up to a maximum per census section.

Regarding the other sources of overrepresentation, these could be easily remedied. On the one hand, to avoid the effect of R1, if the eligible younger persons belong to the same age group (including the quotas of women and men) the choice could be made at random, for example, using the Kish tables (Kish, 1995). On the other hand, to eliminate the attraction effect of 'whole' ages, the solution would be to ask the date of birth instead of the age at last birthday, although later, to reinforce the anonymity of the respondents, only the age at last birthday should be recorded in the corresponding microdata file.

The biases encountered do not invalidate the samples collected by CIS; they simply represent new challenges in research. The same way that analysts try to amend the collected data from the deviations induced by (total or partial) non-response or introduced as a consequence of using quota sampling in the last step (which, as it is well-known,

tends to over-represent the persons that spend more time at home), these new biases should also be taken into account to carry out proper inferences and model analyses. When making inferences, like estimating the voting intention or the climate economic mood of the whole population, the detected biases should be accounted for to correctly weigh the probability that each individual age being included in the sample. Otherwise, given the known inverse relationship between age and ideology in the left-right axis, we might obtain, for example, estimates slightly biased to the left side when making either assessments about global attitudes and values in Spanish society or estimates of trends in public opinion. In our view, however, when modelling, these biases would not introduce additional deviations to the ones already presented in the data[21] once the individual age (not the age group) of each observation was included in the model as an explanatory (predictor) variable. Conditional on age, rules R1 and R2 would have no effect on the collected data, other biases apart. In any case, despite CIS data limitations, it should be noted that the CIS databank comprises the largest and most reliable public database available in Spain for the study of social and political issues. The datasets of CIS can be free downloaded from its website and, as we have already mentioned, its sampling designs has been previously tested with acceptable results by the academia (Stephenson, 1978).

Acknowledgments

We are grateful to Valentín C. Martínez Martín (technical advisor of the CIS research department) for his first-rate assistance kindly answering all our doubts about the CIS sampling designs and to the anonymous referees and an editor for their valuable comments and suggestions. Any inaccuracy is the sole responsibility of the authors. We wish also to thank M. Hodkinson for translation of this paper into English and the support of the Spanish Ministerio de Economía y Competitividad under grants CSO2013-43054-R and ECO2017-87245-R.

References

Bachi, R. (1951). The tendency to round off age returns: measurement and correction. *Bulletin of the International Statistical Institute*, 33, 195–222.

Badal-Valero, E., Alvarez-Jareño, J. A. and Pavía, J. M. (2018). Combining Benford's law and machine learning to detect money laundering. An actual Spanish court case. *Forensic science international*, 282, 24–34. https://doi.org/10.1016/j.forsciint.2017.11.008.

21. As many other survey data, CIS datasets present weaknesses. Their main flaws, however, are not a result of the sampling design followed by CIS but coming from sources of bias (non-response, measurement error, coverage, ...) inherent to the survey approach. Other limitations presented in the data (in particular in the series of data) are, however, imputable to the institution. The historical changes to the board of directives of CIS have often been accompanied by swings in topics of interest, scientific orientation or sense of political opportunity. New issues appear only to disappear soon after or the same issue, just worded differently, is revisited time and time again over the years. This breaks the historical series and makes intertemporal comparisons, at best, difficult. Only a handful of topics (vote intention, government assessment, or political climate, among them) have proven to be resilient.

Brian, A., Baten, J. and Crayen, D. (2009). Quantifying quantitative literacy: Age heaping and the history of human capital. *The Journal of Economic History*, 69, 783–808. https://doi.org/10.1017/S0022050709001120.

Burnap, P., Gibson, R., Sloan, L., Southern, R. and Williams, M. (2016). 140 characters to victory?: Using twitter to predict the UK 2015 general election. *Electoral Studies*, 41, 230–233. https://doi.org/10.1016/j.electstud.2015.11.017.

Carrier, N. (1959). A note on the measurement of digital preference in age recordings. *Journal of the Institute of Actuaries*, 85, 71–85.

Cea D'Ancona, M. A. (2004). *Métodos de encuesta: teoría y práctica, errores y mejora*. Síntesis, Madrid.

Cleveland, W. (1993). *Visualizing Data*. New Jersey: Hobart Press.

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. and Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific Reports*, 7, 40391. https://doi.org/10.1038/srep40391.

Díaz de Rada, V. (2005). *Manual de trabajo de campo en la encuesta*. Cuadernos Metodológicos del CIS, 36. Centro de Investigaciones Sociológicas, Madrid.

Díaz de Rada, V. (2008). La selección de los entrevistados últimos en encuestas presenciales: un análisis de la utilización conjunta del método de rutas y el método de cuotas. *Revista Española de Investigaciones Sociológicas (REIS)*, 123, 209–247. https://doi.org/10.2307/40184898.

Díaz de Rada, V. (2014). Analysis of incidents in face-to-face surveys: Improvements in fieldwork. *Revista Española de Investigaciones Sociológicas (REIS)*, 145, 43–72.

Díaz de Rada, V. (2015). *Manual de Trabajo de Campo en la Encuesta*. Cuadernos Metodológicos del CIS, 36. Centro de Investigaciones Sociológicas, Madrid.

Ebrahimi, M., Yazdavar, H. and Sheth, A. (2017). Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, 32, 70–75. https://doi.org/10.1109/MIS.2017.3711649.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons, New Jersey.

Hope, A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B*, 30, 582–598.

Jungherr, A., Schoen, H. Posegga, O. and Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, 35, 336–356. https://doi.org/10.1177/0894439316631043.

Kalampokis, E., Karamanou, A., Tambouris, E. and Tarabanis, K. (2017). On predicting election results using twitter and linked open data: The case of the UK 2010 election. *Journal of Universal Computer Science*, 23, 280–303.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K. and Gimenez, A. (2016). *Evaluating Online Nonprobability Surveys*. Pew Research Center.

Kish, L. (1995). *Survey Sampling*. John Wiley and Sons, Inc, New York.

Lind, K., Link, M. and Oldendick, R. (2000). A comparison of the accuracy of the last birthday versus the next birthday methods for random selection of household respondents. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 23, 887–889.

Lledó, J., Pavía, J. M. and Morillas, F. G. (2017). Assessing implicit hypothesis in life table construction. *Scandinavian Acturial Journal*, 6, 495–518. https://doi.org/10.1080/03461238.2016.1177585.

Lyons-Amos, M. and Stones, T. (2017). Trends in demographic and health survey data quality: An analysis of age heaping over time in 34 countries in sub saharan africa between 1987 and 2015. *BMC Research Notes*, 10, 1–7. https://doi.org/10.1186/s13104-017-3091-x.

Marsh, C. and Scarbrough, E. (1990). Testing nine hypotheses about quota sampling. *Journal of the Market Research Society*, 32, 485–506.

Mellon, J. and Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research and Politics, July-September*, 1–9. https://doi.org/10.1177/2053168017720008.

Meng, X.-L. (2016). Statistical paradises and paradoxes in big data. In *Royal Statistical Society Annual Conference 2016*. https://www.youtube.com/watch?v=8YLdIDOMEZs.

Mosteller, F., Hyman, H., McCarthy, P., Marks, E. and Truman, D. (1949). *The Pre-Election Polls of 1948*. Social Science Research Council, New York.

Myers, R. (1940). Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America*, 41-2, 395–415.

O'Rourke, D. and Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428–432. https://doi.org/10.2307/3151446.

Pavía, J. M., Badal, E. and García-Cárceles, B. (2016). Spanish exit polls: Sampling error or nonresponse bias? *Revista Internacional de Sociología*, 74, e043. https://doi.org/10.3989/ris.2016.74.3.043.

Pavía, J. M. and García-Cárceles, B. (2012). Una aproximación empírica al error de diseño muestral en las encuestas electorales del CIS. *Metodología de Encuestas*, 14, 45–62.

Pavía, J. M. and Larraz, B. (2012). Nonresponse bias and superpopulation models in electoral polls. *Revista Española de Investigaciones Sociológicas*, 137, 237–264. https://doi.org/10.5477/cis/reis.137.237.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. htpps://www.R-project.org/.

Rodríguez Osuna, J. (1991). *Métodos de Muestreo*. Cuadernos Metodológicos del CIS, 1. Centro de Investigaciones Sociológicas, Madrid.

Rodríguez Osuna, J. (2005). *Métodos de Muestreo. Casos Prácticos*. Cuadernos Metodológicos del CIS, 1. Centro de Investigaciones Sociológicas, Madrid.

Smith, T. (1983). On the validity of inference from non-random samples. *Journal of the Royal Statistical Society, Series A*, 146, 394–403. https://doi.org/10.2307/2981454.

Stephenson, C. B. (1978). A comparison of full-probability and probability-with-quotas sampling techniques in the general social survey. *GSS Technical Report, no. 5. University of Chicago: NORC, Chicago*.

Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

Trochim, W. and Donnelly, J. (2006). *The Research Methods Knowledge Base*. Boston, Massachusetts: Cengage Learning, Boston, Massachusetts.

Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18, 355–366. https://doi.org/10.1080/17457280802305177.

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, 15, 335–350.

Yang, K. and Banamah, A. (2014). Quota sampling as an alternative to probability sampling? an experimental study. *Sociological Research Online*, 19, 1–11. htpps://doi.org/10.5153/sro.3199.

Selected articles from
XVI Conferencia Española
de Biometría 2017

SORT 42 (2) July-December 2018, 209-236 DOI: 10.2436/20.8080.02.75

Effect of agro-climatic conditions on near infrared spectra of extra virgin olive oils*

M. I. Sánchez-Rodríguez*,[1], E. M. Sánchez-López[2], J. M. Caridad[1], A. Marinas[2] and F. J. Urbano[2]

Abstract

Authentication of extra virgin olive oil requires fast and cost-effective analytical procedures, such as near infrared spectroscopy. Multivariate analysis and chemometrics have been successfully applied in several papers to gather qualitative and quantitative information of extra virgin olive oils from near infrared spectra. Moreover, there are many examples in the literature analysing the effect of agro-climatic conditions on food content, in general, and in olive oil components, in particular. But the majority of these studies considered a factor, a non-numerical variable, containing this meteorological information. The present work uses all the agro-climatic data with the aim of highlighting the linear relationships between them and the near infrared spectra. The study begins with a graphical motivation, continues with a bivariate analysis and, finally, applies redundancy analysis to extend and confirm the previous conclusions.

MSC: 62H20, 62Pxx, 82-08, 62-09

Keywords: Extra virgin olive oil, infrared spectroscopy, agro-climatic data, linear correlations, redundancy analysis

1. Introduction

Spain is the first worldwide producer of extra virgin olive oil (EVOO), where Andalusia encompasses 80% of the national production. EVOO is an edible oil very much appreciated by its flavour and benefits for health. Its high quality could be affected by frauds in marketing, such as adulteration with other cheaper oils (for example, palm, corn, hazelnut or refined olive oil) or with the indication of a false geographical origin. These practices considerably modify its quality indexes. Therefore, authentication of EVOO requires fast, reliable and cost-effective analytical procedures which require no or little sample manipulation, such as near infrared spectroscopy (NIR). Contrary to classical separation techniques (for example, gas chromatography), NIR spectra provide

* *Corresponding author:* td1sarom@uco.es. Avda. Puerta Nueva, s/n. 14071. Córdoba

[1] Department of Statistics and Business. University of Cordova.

[2] Department of Organic Chemistry. University of Cordova.

Received: March 2018
Accepted: October 2018

continuous information rich in both isolated and overlapping bands and their analysis requires the application of multivariate statistics (see Öztürk, Yalçin and Özdemir, 2010).

There are in the literature many examples of the application of chemometrics to determine qualitative and quantitative information of EVOO from NIR spectra, specially, with the aim of its authentication. For instance, Bertran et al. (2000) apply NIR and pattern recognition as screening methods for the authentication of EVOO of very close geographical origins. Mailer (2004) shows a rapid evaluation of olive oil quality by NIR reflectance spectroscopy. Galtier et al. (2007) determine geographic origins and compositions of EVOO by chemometric analysis of NIR spectra. Woodcock, Downey and O'Donnell (2008) show a confirmation of declared provenance of European EVOO samples by NIR spectroscopy. Casale et al. (2012) present a characterization of Protected Designations of Origin (PDO) olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR (mid-infrared) spectroscopy) and selective (fatty acid composition) analytical techniques. Finally, some previous papers of our research group (see Sánchez-Rodríguez et al. (2013) and Sánchez-Rodríguez et al. (2014)) show new chemometric approaches to empathize the potential of NIR and MIR spectra to determine the fatty acid profile of EVOO, the fatty acids being its major components and considered as a quality parameter in order to its authentication. Therefore, NIR and MIR spectra contain valuable and diverse information about EVOO.

Moreover, there are in the literature many works analysing the influence of weather, agro-climatic or meteorological[1] conditions on food content, in general, or in EVOO components, in particular. Thus, for example, Martínez-Herrera et al. (2006) analyse the chemical composition of *Jatropha curcas L.*, a multipurpose shrub of significant economic importance because of its several potential industrial and medicinal uses, from different agro-climatic regions of Mexico. Jarvis et al. (2008) and Khokhar et al. (2017) study the influence of agro-climatic conditions on wheat in western Canada and India, respectively. Zheng et al. (2012) show the effects of latitude and weather conditions on the contents of black currant, while Yang et al. (2017) analyse the same effects on Finnish berries. Falasca, Ulberich and Ulberich (2012) develop an agro-climatic zoning model to determine potential production areas for castor bean. Luciano et al. (2013) treat the effects of the weather and the soil on the composition of grapes. Rymbai et al. (2014) study the physiological characteristics of mango in different agro-climatic regions of India. Edmunds et al. (2015) analyse the relationships of preharvest weather conditions and soil factors to susceptibility of sweetpotato. Dorey et al. (2016) model sugar content of pineapple under agro-climatic conditions on Reunion Island. Finally, there are many papers treating the effect of weather and agro-climatic conditions on oils (such as Leskinen, Suomela and Kallio, 2009a and Leskinen et al., 2009b), especially the numerous studies of olive oils: for example, Sacco et al. (2000), D'Imperio et al.

1. Climatology deals with the scientific study of climate, that is, the processes and phenomena of the atmosphere over relatively long periods of time. However, Meteorology studies the characteristics of the atmosphere over a short period of time, especially as a means of forecasting the weather. The agro-prefix placed before both terms refers to the interrelationship between Climatology and Meteorology with the processes of agricultural production.

(2007), Cornejo, Bueno and Gines (2012), Awan (2014), Alowaiesh, Singh and Kailis (2016), Ozdemir (2016), Veizi, Peçi and Lazaj (2016), Zaied and Zouabi (2016) and Merchak et al. (2017). But there are few studies considering NIR data to study this agro-climatic influence on oils or other food products.

Regarding the multivariate statistical technique being applied, the majority of the studies included in the literature consider a single factor, a non-numerical variable, to establish different meteorological or agro-climatic zones – see, for example, Alowaiesh et al. (2016), Cornejo et al. (2012), Leskinen et al. (2009a) and Leskinen et al. (2009b), Merchak et al. (2017) or Zheng et al. (2012). If this factor is used as an independent variable in a statistical model, ANOVA (or MANOVA) and a post-hoc test can be used to compare the means corresponding to the defined zones in a numeric variable. The agro- climatic factor can also be used as a dependent variable in the linear discriminant analysis (LDA), where the high dimensionality of the independent variables can be reduced by previously applying principal component analysis (PCA) or partial least squares (PLS). However, the present study rather uses the complete agro- climatic data base obtained from the official webpage of the Automatic Weather Stations (AWEs) of Andalusia. In particular, the historical daily information from 2005 to 2010 has been downloaded for the following variables: temperature, humidity, wind speed, radiation, precipitation and evapotranspiration.

In this case, the agro-climatic data are aggregated in different ways and associated to the EVOO (taking into account the nearest AWE) by using computational programs designed by the powerful free software R-project (R Core Team (2018)). The aim of the study is to explore the linear relationships between agro-climatic and EVOO NIR data: firstly, by using bivariate correlation analysis and, then, generalizing the procedures to multivariate analysis with the application of Redundancy Analysis (RDA).

In particular, Section 2 describes the process of acquisition of NIR and agro-climatic data, the statistical bivariate and multivariate methodology and the computational implementation. Section 3 shows the results and discussion: firstly, the graphical analysis of NIR (original and derivative) spectra and the series of agro-climatic data; secondly, the results of the correlation analysis between the agro-climatic measurements and the spectral absorbance are shown; thirdly, some of the previous conclusions are confirmed and extended by the application of the multivariate technique of RDA. Finally, Section 4 includes the main conclusions of the work.

2. Materials and methods

2.1. Data

2.1.1. NIR data

Olive oil was extracted by the producers through a two-phase centrifugation system. Information from 222 Andalusian extra virgin olive oils, collected from consecutive

harvests from 2005-06 to 2010-11 (denoted H1, H2,..., H6, respectively), is available. The chemical data from each EVOO have been provided by near-infrared (NIR) spectroscopy by the staff of the Organic Chemistry Department of the University of Cordova (Spain). The instruments employed for spectra collection were available at Central Service of Analyses (SCAI) and included a Spectrum One NTS FT-NIR spectrophotometer (Perkin Elmer LLC, Shelton, USA) equipped with an integrating sphere module. Samples were analysed by transflectance by using a glass petri dish and a hexagonal reflector with a total transflectance pathlength of approximately 0.5 mm. A diffuse reflecting stainless steel surface placed at the bottom of the cup reflected the radiation back through the sample to the reflectance detector. The spectra were collected by using Spectrum Software 5.0.1 (Perkin Elmer LLC, Shelton, USA). The reflectance (log 1/R) spectra were collected with two different reflectors. Data correspond to the average of results with both reflectors, thus ruling out the influence of them on variability of the obtained results. Moreover, spectra were subsequently smoothed using the Savitzky-Golay technique, which performs a local polynomial least squares regression in order to reduce the random noise of the instrumental signal (Savitzky and Golay (1964)). Once pre-treated, NIR data of 1237 measurements for each case (representing energy absorbed by olive the oil sample at 1237 different wavelengths, from 800.62 to 2499.64 nm) were supplied to the Department of Statistics (University of Cordova) in order to be analysed (Figure 1).

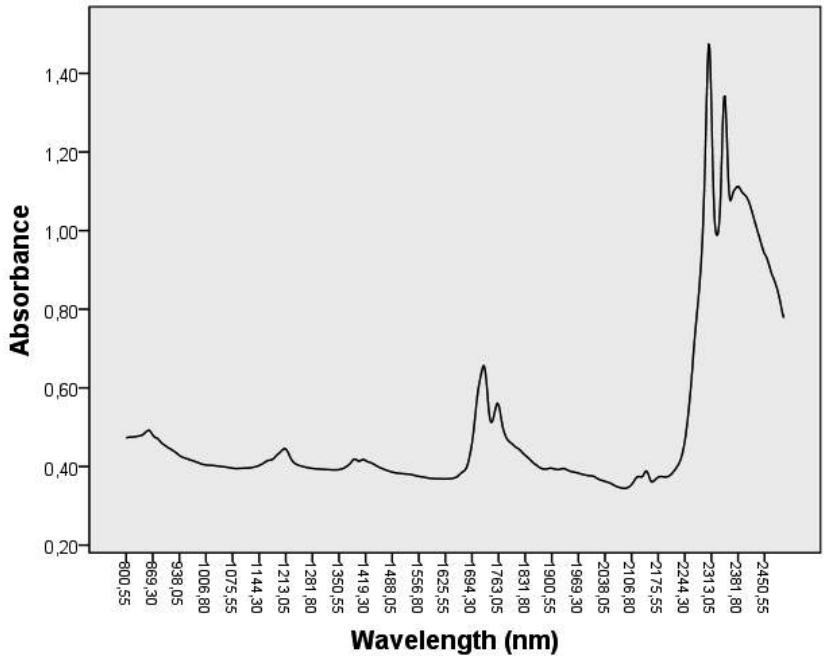


Figure 1: *NIR spectrum of an extra virgin olive oil.*

2.1.2. Agro-climatic data

The agro-climatic data used in the work has been obtained of the official website of the Andalusian Institute of Agricultural, Fisheries, Agrifood and Organic Production Research and Training (IFAPA). In this webpage, the long-run information registered in the Automatic Weather Stations (AWEs) can be accessed[2]. These stations have a suitable plan of maintenance and an exhaustive review of the records that supply the sensors. There are approximately 120 AWEs in all the Andalusian provinces, though in this work only the historical daily information corresponding to the 28 AWEs specified in Table 1 (see Appendix A), for the period 2005-2010 (years before the considered harvest years), has been downloaded. These AWEs have been selected due to their proximity with the point of extraction of the available oils.

Information about the following variables has been considered in this study:

- *Temp*: Daily average temperature, in °C. The temperature is measured by a sensor Pt1000 whose functioning is based on the variation of the resistance of the platinum element by the temperature.
- *Hum*: Daily average relative humidity, in %. The measurement of the relative humidity is realized by a capacitive device of solid condition: sensor HUMICAP 180, plastic polymer that tends to absorb humidity. The sensor changes its electrical characteristics by the variations of humidity, in such a way that diminishes its electrical capacity by the absorption of dampness.
- *WSpe*: Daily average wind speed, in meters per second. Its measurement is realized by a weather vane, in which the rotation of a propeller produces an electrical sign in alternating current, of frequency proportional to the wind speed.
- *Rad*: Daily average radiation, in MJ per m^2. The measurement is realized by a pyrometer constituted by a photoelectric cell of silicon being sensitive to the radiation from 350 to 1100 nm, orientated in a southerly direction and ensuring that another sensor or accessory of the tripod does not cast shade on it.
- *Precip*: Daily precipitation, in mm. The AWE has a device of swinging small containers to measure the volume of rainfall, that is measured by the number of contacts with a tab of the device (each one equivalent to 0.20 mm) that are produced by the overturning of the rain water from one container to the other.
- ET_0. The potential evapotranspiration (PET) is the loss of dampness (in mm per day) of a surface for direct evaporation together with the water loss for perspiration of the vegetation. PET represents the maximum quantity of water that can evaporate from a soil completely covered with vegetation, which develops in ideal conditions and supposing that there are no limitations in the availability of water. ET_0, denoted here as ETo for purposes of labelling, is similar to the ETP though

2. The link is https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController, where the historical data can be downloaded by clicking on the name of the station and selecting the agro-climatic measurements and the start and end dates [accessed on 02 October 2018].

it is applied to a specific or standard cultivation, habitually cereals or alfalfa, from 8 to 15 cm of uniform height, of active growth, totally covering the soil and not being submitted to water deficit.

2.2. Methodology

2.2.1. Bivariate analysis

Pearson's linear correlation coefficient, r, determines the degree of linear association existing between two numerical variables, being higher as the coefficient is closer to 1 in absolute value. Assuming bivariate normality of the variables, and under the null hypothesis of zero correlation, the statistic $t = r\sqrt{(n-2)/(1-r^2)}$ has t-Student distribution with $n-2$ degrees of freedom, where n is the sample size, equal to 222 in this study. Using a significance level of $\alpha = 0.05$, values of r such as $-0.1317 < r < 0.1317$ show no statistical evidence for rejecting the hyothesis of zero correlation.

2.2.2. Redundancy analysis

Canonical redundancy analysis (RDA) and canonical correspondence analysis (CCA) are two forms of asymmetric canonical analysis, where asymmetric means that the matrices involved in the analysis, $\mathbf{X}$ and $\mathbf{Y}$, do not play the same role: $\mathbf{Y}$ is a matrix of response variables – in this case, containing the spectral information – and $\mathbf{X}$ is the matrix of explanatory variables – in this study, the agro-climatic measurements. This aspect contrasts with canonical correlation analysis where the two matrices play the same role in the analysis and so can be interchanged. $\mathbf{X}$ is used to explain the variation in $\mathbf{Y}$, as in regression analysis, in two steps[3] [4]:

1. Multivariate regression of $\mathbf{Y}$ on $\mathbf{X}$, which is equivalent to a series of multiple linear regressions of the individuals variables of $\mathbf{Y}$ on $\mathbf{X}$ and produces a matrix of fitted values $\hat{\mathbf{Y}}$.
2. Principal component analysis (PCA) of $\hat{\mathbf{Y}}$ in order to reduce its dimension. PCA components of $\hat{\mathbf{Y}}$, called RDA components or *redundancy axes*, are obtained as a reduced number of linear combinations of the variables of $\hat{\mathbf{Y}}$, orthogonal among themselves, explaining a maximum percentage of their variability.

Therefore, in RDA the variability of the variables of $\mathbf{Y}$ are explained from PCA components (factors or latent variables) depending on the variables of $\mathbf{X}$ and so RDA can be seen as a constrained version of PCA.

3. $\mathbf{X}$ and $\mathbf{Y}$ are generally standardized to eliminate the effect of the measurement units.

4. The main assumptions of the data are linearity between the variables of matrix $\mathbf{Y}$ and the variables of the matrix $\mathbf{X}$ and the variance homogeneity of each set of data.

Each eigenvalue of the correlation matrix of the variables of $\hat{\mathbf{Y}}$, λ_j for $j = 1, \ldots, g$, represents the variance of each redundancy axis, whose direction is calculated from the corresponding eigenvector. The proportion of the total variance of $\mathbf{Y}$ explained by a redundancy axis k, $k = 1, \ldots, g$, is given by:

$$\frac{\lambda_k}{\sum_{j=1}^{g} \lambda_j}.$$

The *redundancy index* of the model (similar to a coefficient of determination) is defined by:

$$R_m^2 = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{g} \lambda_j},$$

being m the number of redundancy axes (among the possible g RDA components) to retain.

The results of the applications of RDA analysis are usually shown by representing both matrices, $\mathbf{X}$ and $\mathbf{Y}$, in a space of reduced dimension: the two or three-dimensional space formed by the first RDA components. Variables or cases with the highest coordinates (scores) in a RDA component or redundancy axis are very useful to interpret it, showing the variables and/or cases that are discriminated by this RDA component. Besides, the proximity between variables, cases or RDA components represents the high association between them.

Redundancy analysis as an alternative for canonical correlation analysis was presented by authors such as Rao (1964) and van den Wollenberg (1977). More recently, Legendre, Oksanen and ter Braak (2011), test the significance of the redundancy axes in RDA.

2.2.3. Functional data analysis

For some years, the computing applied to different areas has caused a major technological change due to the addition of faster and more precise measuring equipments. This fact affects one of the paradigms on classical statistics: the number of data should be greater than the number of variables. Currently, large databases corresponding to observations of random variables taken over a continuous interval (or increasingly extensive discretizations of this continuous interval). This kind of data, named *functional data*, appear in a natural way in fields such as the spectrometry, where the measurement result is a curve, a spectrum (see, for example, Aguilera et al. (2010) or Saeys, De Ketelaere and Darius, 2008).

Moreover, in chemometrics, the treatment of a spectrum in the context of functional data analysis, as a continuous function, enables the obtaining of the spectral derivatives as any differentiable function must be continuous at every point in its domain. Many studies of different fields, in particular, of olive oil have proven that the first or second derivative of NIR spectra provide valuable qualitative or quantitative information about

oil that, however, the original spectra do not show (see, for example, Chen et al. (2015) or Woodcock et al. (2008)). Although the original spectral curves overlap, sometimes those ones associated to a high content in a concrete compound or having the effect of an external factor show show higher variability. Therefore, these variations or discrepancies are appreciated more clearly in the first derivative of the spectra than in the original spectra.

2.2.4. Computational implementation

The agro-climatic data corresponding to the year previous to the olive harvest and to the nearest AWE (or the average of the nearest AWEs) are associated to each oil sample. A procedure has been programmed, using R, that permits to select the considered agro-climatic variable (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *ETo*) and accumulates the daily measurements corresponding to several days or months. In particular, the following function has been defined:

AGR-CLIM-function(*station*, *harvest*, *month1*, *month2*, *agro-climatic measurement*),

with the following arguments:

- *station*: among the 28 observed AWEs, the case has associated the code of the nearest geographically (see Table 1),
- *harvest* years, from 1 (2005-06) to 6 (2010-11),
- given the station and the harvest, the period of time (from *month1* to *month2*) can be selected to aggregate the daily agro-climatic measurements,
- *agro-climatic measurement*, distinguishing among the 6 previously described: *Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *ETo*,

The function returns as value the aggregated agro-climatic measurement according to the selected months and the established meteorological criterion.

Having extracted the data, Pearson's linear correlation is computed between the different agro-climatic measurement, aggregated for different months, and some spectral values of absorbance for the original spectra or their (first or second) derivatives. The graphical procedures will mark in all cases the correlation coefficients which are (or not) statistically different from zero (with $\alpha = 0.05$). As stated above, for the sample size $n = 222$ they are the values outside the range $(-0.1317, 0.1317)$.

The packages of R-project 'fda' (Ramsay et al. (2017)) and 'fda.usc' (Febrero-Bande and Oviedo de la Fuente, 2012) have been used to obtain the spectral derivatives and the multivariate analysis of RDA has been developed by using the package 'vegan' (Oksanen et al. (2018)). Detailed information of the code of the programs designed to read the agro-climatic and chemical data, including the above-mentioned function, and to obtain the diverse range of graphics considered in the study can be seen in the Supplementary Material.

3. Results and discussion

3.1. Graphical analysis

3.1.1. Analysis of NIR spectra

NIR spectra are the representation of the *absorbance*, that is, the quantity of energy absorbed by an oil at each wavelength (from 800.62 to 2499.64 nm, 1237 measures in total). As indicated above, the continuous treatment of a spectrum, instead of an extensive discretization, permits the obtaining of its derivatives that, in occasions, contain valuable information about olive oil compositions.

Thus, in Figure 2 the original spectra as well as their two first derivatives are represented, where the spectra are grouped in the same colour corresponding to a same harvest. The visual analysis highlights the separation or divergence of some spectra, especially those corresponding to the last harvest (H6, depicted in pink). This discrepancy is more pronounced in some ranges of wavelengths of derivative spectra, whose detail is represented in Figure 3 (where the points of maximum discrepancy are denoted by P_1, $P_2, \ldots, P_{10}$ for future analysis).

In addition, in Figure 4 the transposes of the original spectra and their two first derivatives are shown, i.e., the curves are represented as a function of the case. This graphic also highlights the structural change corresponding to the last harvest, H6; this change is especially evident by the view of the derivative spectra.

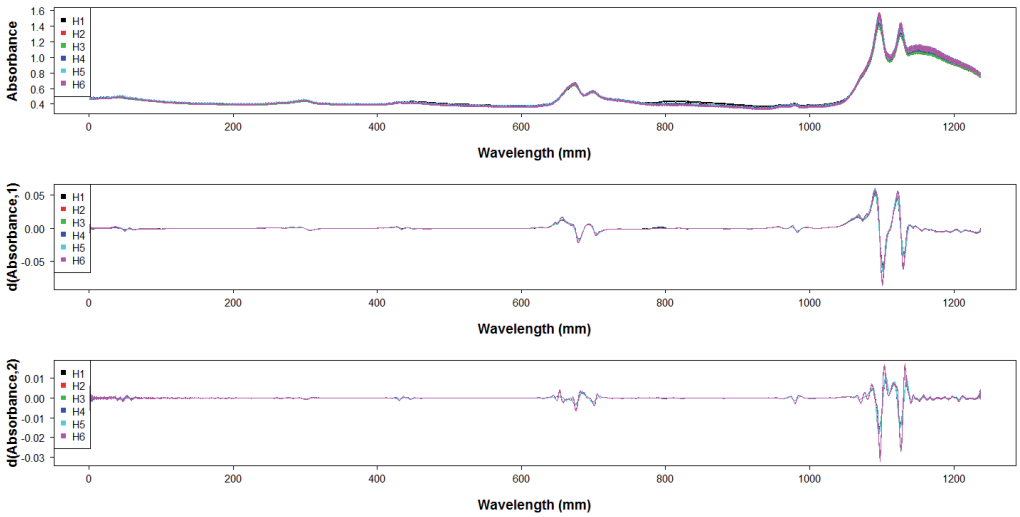


Figure 2: *NIR spectra and their first and second derivatives.*

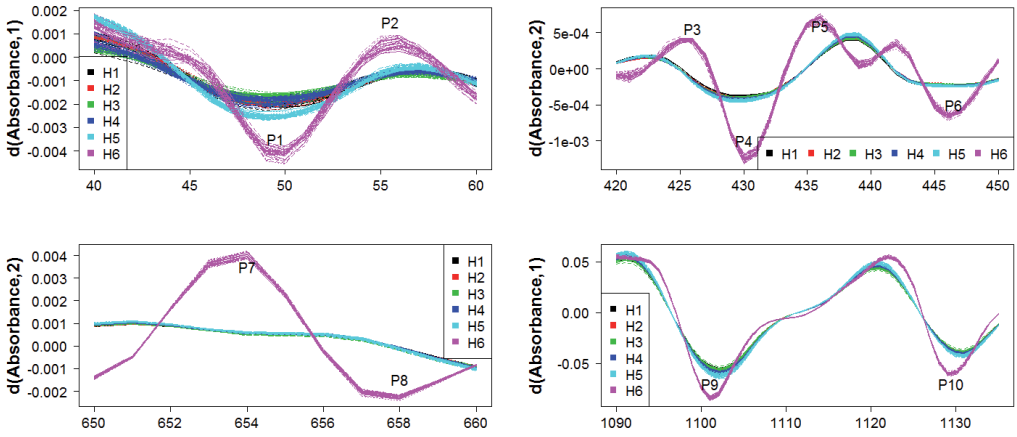


Figure 3: *Spectral details corresponding to the maximum discrepancies.*

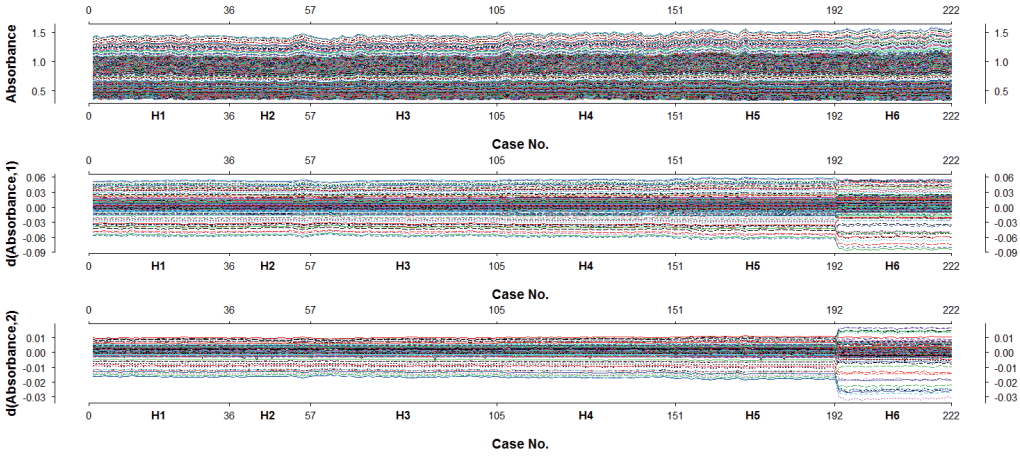


Figure 4: *Spectral and derivative NIR values, as a function of the case.*

3.1.2. Analysis of series of agro-climatic data

In this section, the series of the six agro-climatic measurements (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *Eto*) are represented for the six harvests. The daily values are accumulated for each month (afterwards, the reason is explained) and then standardized in order to eliminate the effect of the measurement units of each variable. So, dimensionless series are obtained that can be represented and compared in the same graphic. These standardized values are represented in Figure 5 which shows a cyclical tendency for all the considered variables. In general, the proximity of the trajectories of evolution of the variables *Temp*, *WSpe*, *Rad* and *Eto*, on the one hand, and *Hum* and *Precip*, on the other, is observed, noting also the symmetry among them. With regard to the relation between precipitation and radiation, Bradley et al. (2011) use cross-spectral analysis

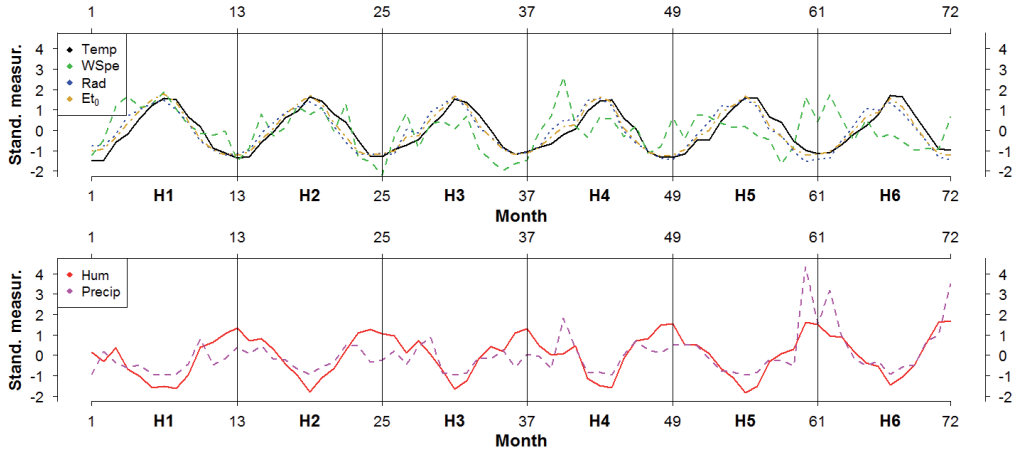
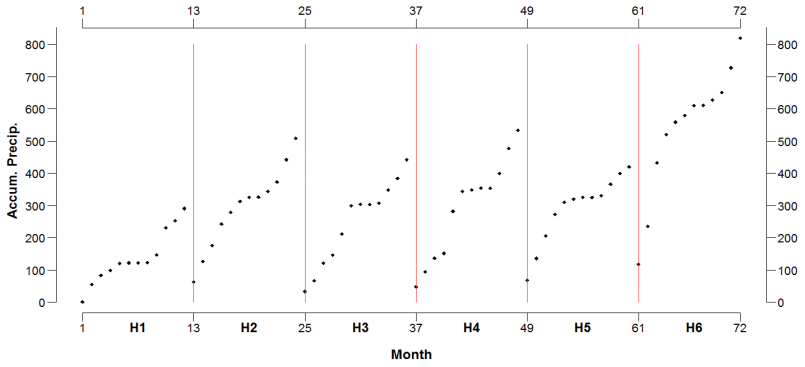


Figure 5: *Monthly accumulated and standardized agro-climatic measurements.*



Year	Ann. prec. (in mm.)
2005	290.65
2006	508.97
2007	442.13
2008	533.80
2009	419.85
2010	819.05

Figure 6: *Monthly accumulated precipitation (in mm.), for each harvest.*

to show that precipitation has a role to play in the maintenance of phenology cycles because it maintains constant vegetation growth reducing so the seasonal impact of the solar radiation.

As fundamental irregularity of Figure 5, the especially high values of the variables that represent the wind speed (*WSPe*, in green) and the volume of precipitation (*Precip*, in pink) at the beginning of the 4th harvest and at the beginning and the end of the 6th harvest (H6) can be highlighted. This fact corroborated the work of Back and Bretherton (2005) which studied the relationship between wind speed and precipitation in the Pacific and found a significant correlation between these variables. The specially irregular behaviour of the *Precip* variable in H6, whose accumulated mean values are specially high, can also be deduced from the observation of Figure 6.

Therefore, the anomalous accumulated precipitation (or wind speed) values corresponding to the sixth harvest together with the anomalous derivative NIR spectra corresponding to the same harvest justify the formulation of the following question: What

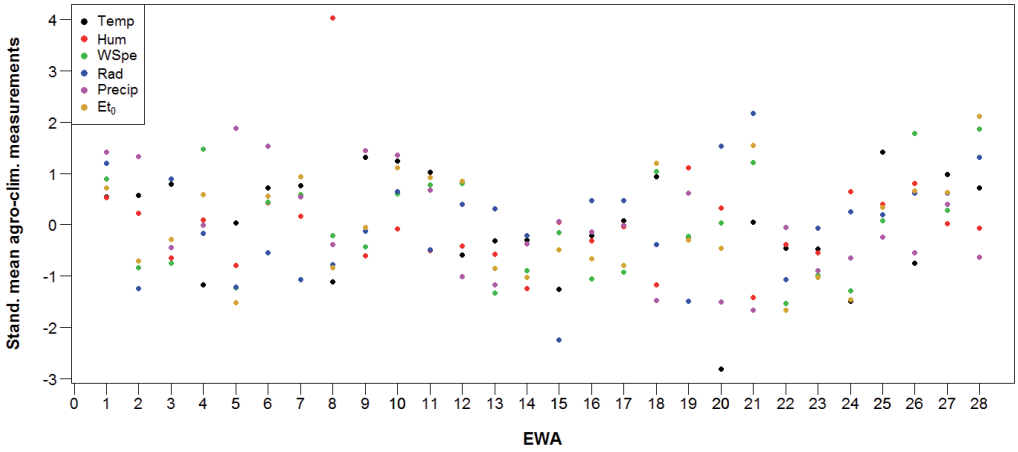


Figure 7: *Standardized mean agro-climatic measurements, for each AWE.*

is the effect of the precipitation or the wind speed, in particular, or the agro-climatic conditions, in general, on NIR spectra or on the chemical compounds of EVOO?

Finally, Figure 7 depicts the standardized mean values for the six agro-climatic measurements for the 28 automatic weather stations. The obvious discrepancies among the mean values corresponding to the different AWEs makes reasonable the assignation the agro-climatic measurements associated to the nearest AWE to each olive oil (case).

3.2. Bivariate analysis

The following function:

$$\textit{AGR-CLIM-function}(\textit{station}, \textit{harvest}, \textit{month1}, \textit{month2}, \textit{agro-climatic measurement}),$$

described in Section 2.2 (Methodology) and whose code is included in the Supplementary Material, has been applied to each EVOO (222, in total), considering the nearest AWE (station) and the corresponding harvest. The six agro-climatic measurements previously downloaded (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip* and *Eto*) have been accumulated for each month, from January to December. Therefore, a list of 12 matrices of dimension 222×6, $[\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_{12}]$, is available. Moreover, $\mathbf{Y}$ is the matrix of dimension 222×10 whose columns contain the absorbance associated to the 10 peaks of maximum discrepancy (P_1, P_2, ..., P_{10}) represented in Figure 3.

The aim of aggregating the agro-climatic measurements has been to relate them more adequately to the phenological cycle of the olive grove, which will directly influence the composition of the oil. As shown in Figure 8, this cycle is not distributed equally, and in this way the months of interest in each case could be studied independently. In the bibliography, authors such as Orlandi et al. (2012), in the study of the influence of

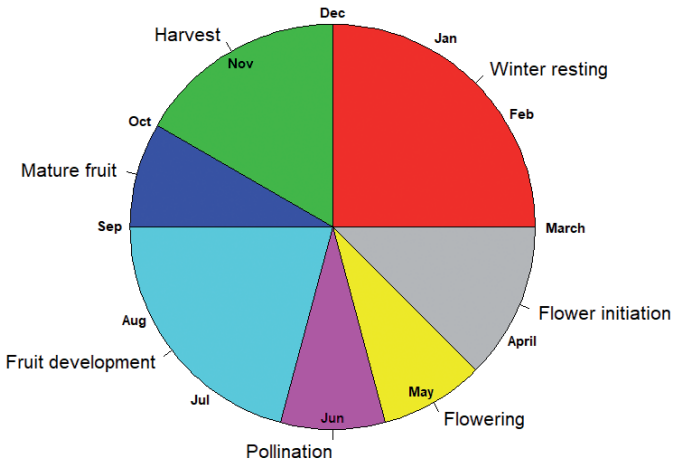


Figure 8: *Phenological stages of olive.*

climate data on oil production in southern Italy, also consider meteorological variables on a monthly basis.

3.2.1. Correlations between the agro-climatic measurements and the discrepancy spectral peaks

In this section, Pearson's linear correlation coefficients are calculated between each of the six agro-climatic measurements, accumulated for each month, and the discrepancy spectral peaks denoted in Figure 3. The results are shown in Figures B.1-B.4 (in Appendix B), where the light grey lines of points mark the correlations -0.5 and 0.5 and the dark grey lines of points show the frontier between the values being different (or not) statistically from zero for $\alpha = 0.05$.

The following fundamental conclusions can be deduced from the observation of Figures B.1-B.4:

- There are many high correlations, next to -1 or 1, specially for the accumulated agro-climatic measurements corresponding to January, February, March, June and November. Therefore, the lowest correlations between the discrepancy spectral peaks and the aggregate agro-climatic measurements appear for the months of the phenological stages corresponding to the development and the maturation of olives (see Figure 8). And so it may be interpreted that the highest effect of the meteorological conditions (in particular, of the precipitation), reflected in NIR spectra, takes place not on the fruit but on the tree.
- From the observation of the different agro-climatic measurements, the precipitation (*Precip*, in pink) and the radiation (*Rad*, in blue) are the variables showing, in general, the highest (positive or negative) correlations, having opposite sign.

As shown in Bradley et al. (2011), precipitation and radiation have negative linear correlation and now Figures B.1-B.4 highlight that both agro-climatic measurements have a contrary effect on the discrepancy spectral peaks. Besides, the sign of the pairwise correlations between *Precip-Rad* and the peaks P_2, P_3, P_5, P_7 are the same, and the opposite of the sign of the correlations between the rest of the peaks. By coincidence, these peaks are the relative maxima of the derivative NIR spectra while the other peaks are the relative minima.

- Some agro-climatic variables are almost uncorrelated between the discrepancy spectral peaks for many months but, nevertheless, shown values closer to 1 (in absolute terms) for a concrete month. These are the case, for example, of the evapotranspiration (*ETo*, in yellow) or the humidity (*Hum*, in red) in March or November, whose influence on the spectral peaks is the contrary. The negative or inverse correlation between both variables can be intuited from the observation of Figure 5. Besides, in March and November, the standardized values for *ETo* and *Hum* are quite similar and, however, the effect on the discrepancy spectral peaks is the highest.

3.2.2. Correlations between the agro-climatic measurements and the spectral absorbance

In this section, Pearson's linear correlation coefficients between the monthly accumulated agro-climatic measurements and the spectral absorbance are calculated. The results, that coincide with the ones obtained from Figures B.1-B.4, are shown in Figures C.1-C.4 (in Appendix C).

The following general conclusions can be obtained:

- January and December are the months showing, in general, the highest correlations (in absolute terms) and April is the one with the correlation values nearest to zero. This fact confirms, newly, that the highest correlations appear in the phenological stage of winter resting of olive tree (see Figure 8).
- Taking into account the different agro-climatic measurements, the precipitation (*Precip*, in pink) and the radiation (*Rad*, in blue) are the variables showing the highest correlations, being the opposite the sign of their linear correlation. In general, the sign of the correlation for the radiation and the evapotranspiration (*ETo*, in yellow) is the same, and the opposite to the sign of the correlation for all other variables.

3.3. Multivariate analysis

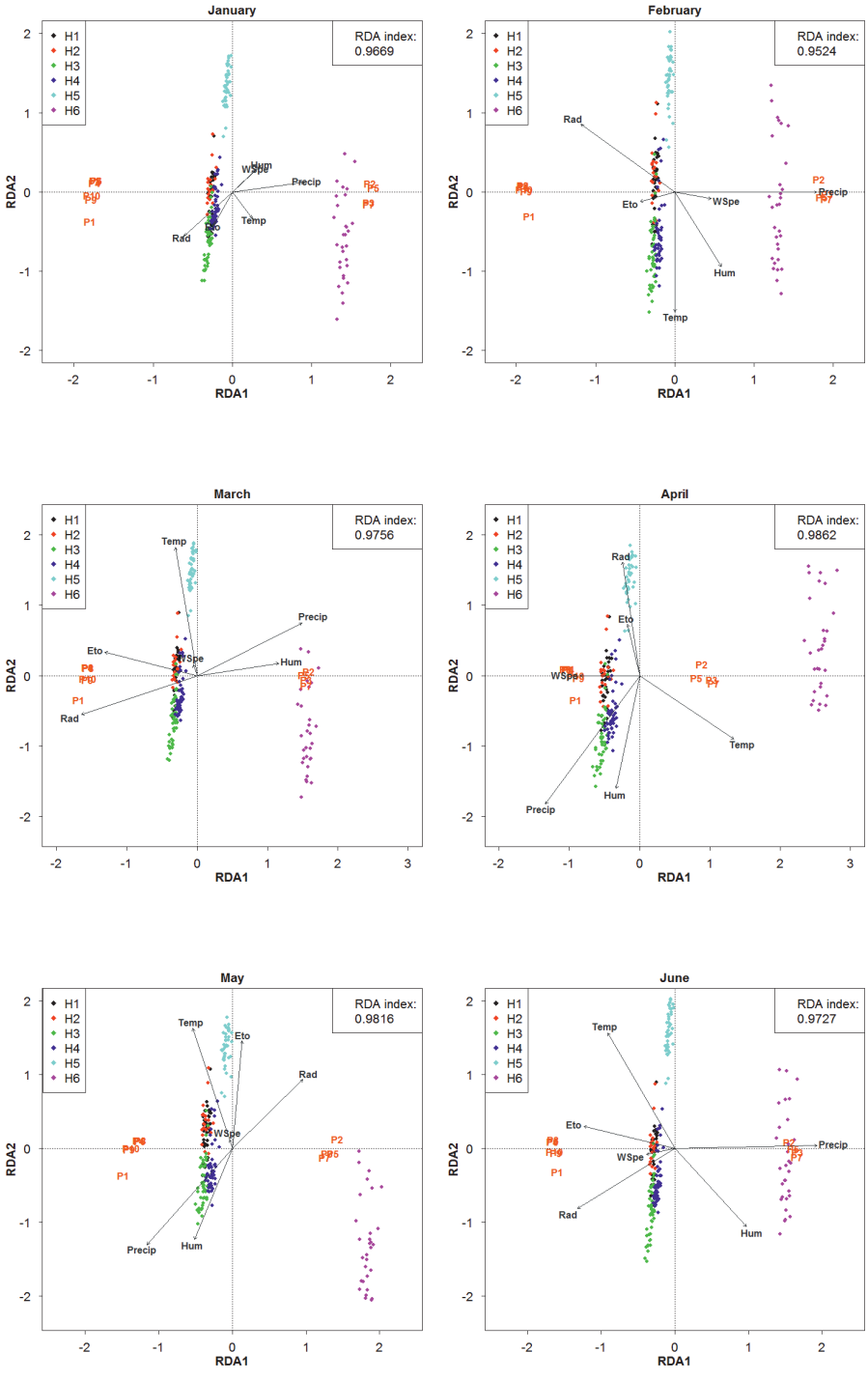
In this section, redundancy analysis (RDA) is applied to generalize the previous results and highlight the cause and effect relationships between two data matrices: one of them, the matrix of explanatory variables, containing in its columns the six accumulated agro-climatic measurements for a specific month ($\mathbf{X}_i$, $i = 1, \ldots, 12$) and the other, the matrix

of response variables ($\mathbf{Y}$), formed by the spectral absorbance associated to the 10 peaks of maximum discrepancy (P_1,P_2,...,P_{12}) represented in Figure 3.

The results of the application of each RDA are shown in the two-dimensional space formed by the two first RDA components (RDA1 and RDA2), where both matrices, $\mathbf{X}_i$ and $\mathbf{Y}$, are represented. The results are drawn, for each month, in Figure 9. Each individual representation shows: *a)* the cases, using different colors for the different harvests (black, red, green, blue, cyan and pink for H1,H2,...,H6, respectively); *b)* the response variables: the absorbance for the spectral peaks ($\mathbf{Y}$, in orange); *c)* the explanatory variables: the agro-climatic measurements for each month ($\mathbf{X}_i$, $i = 1,\ldots,12$, in gray). The redundancy index is greater than 0.95 for all the months (as it is shown at the top right of each graphic), which indicates that the percentage of the total variance of $\mathbf{Y}$ (spectral peaks) explained by the two first RDA components is greater than 95%.

In general lines, the conclusions obtained from the observation of Figure 9 confirm some of the above-mentioned ones, deduced from the bivariate analysis. More in particular, the following results can be enumerated:

- *Cases analysis*: The cases corresponding to the last harvest (H6, in pink) are clearly discriminated or separated from the remaining harvests for all the months: cases of H6 have high scores (in absolute terms) in RDA1 for all the months whereas all the other cases have scores near zero in this axis. RDA2 permits to discriminate the harvest H5 (in cyan) from the others: cases of H5 have high (absolute) scores in RDA2. Cases of H6 have also high scores in RDA2 for months such as October but the groups of cases can be discriminated by the scores in RDA1. The cases associated to H1, H2, H3 and H4 are, in general, overlapped and, so, they are not discriminated by RDA1 and RDA2 (the most important redundancy axes), showing scores near zero in both redundancy axis, in general. RDA statistically modelled the situation previously represented in Figure 3, where spectra corresponding to H6 (and H5, to a lesser degree) are clearly discriminated from the others for some ranges of the original spectra or their first two derivatives.
- *Response variables analysis*: P_2, P_3, P_5, P_7 are clearly discriminated from the other spectral peaks in all the months (being all of them depicted in orange). For all months, the peaks have scores greater than one, in absolute terms, in RDA1, whereas the scores in RDA2 are near zero. In this case, RDA has also a clear correspondence with the representation of Figure 3, as P_2, P_3, P_5, P_7 are relative maxima of the derivative spectra while P_1, P_4, P_6, P_8, P_9 and P_{10} are relative minima.
- *Explanatory variables analysis*: As in bivariate analysis, taking into account the different agro-climatic measurements (represented in gray), the radiation (*Rad*) and the precipitation (*Precip*) are the variables having the highest scores (in absolute terms) in RDA1 (especially, in February, June and December). The sign of the scores (and, so, the correlation) is the opposite for these two variables (in line with the observation of Figure 5 and Bradley et al. (2011)). The humidity (*Hum*)


January
RDA index: 0.9669
February
RDA index: 0.9524
March
RDA index: 0.9756
April
RDA index: 0.9862
May
RDA index: 0.9816
June
RDA index: 0.9727
H1
H2
H3
H4
H5
H6
RDA1
RDA2
Temp
Precip
Rad
Eto
WSpe
Hum

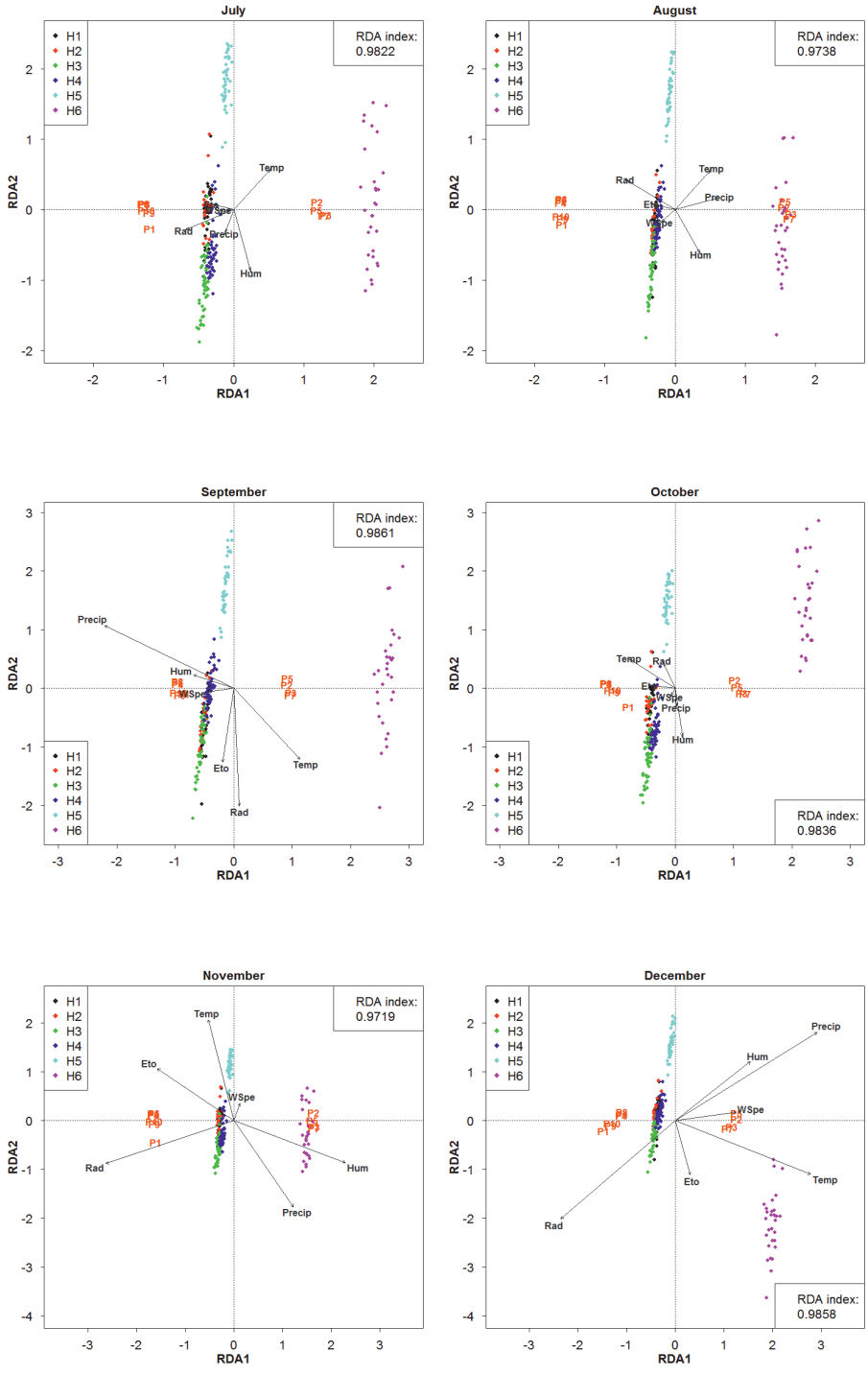


Figure 9: *RDA representations from January to December.*

and the temperature (*Temp*) are clearly discriminated by RDA2 for months such as May or June, having both agro-climatic measurements opposite scores. The evolution of these two variables is also the opposite in Figure 5. Finally, with respect to the cases, response and explanatory variables, in months such as January, March or November, *Precip* shows high scores, in absolute terms, in RDA1 close to the scores of the relative maxima peaks P_2, P_3, P_5, P_7 and the last harvest (H6). Besides, in March, May or June, *Temp* shows high (absolute) scores in RDA2, this variable being near the cases of H5.

4. Conclusions

During recent years NIR spectroscopy has been commonly used because it is a fast, reliable and cost-effective chemical technique. Many studies apply chemometrics analysis to highlight the valuable information contained in NIR spectra of EVOO. Firstly, studies such as Galtier et al. (2007) or Sánchez-Rodríguez et al. (2013) and Sánchez-Rodríguez et al. (2014) show the prediction of the fatty acid profile (quantitative information) from NIR spectra. Other authors (Bertran et al. (2000) or Öztürk et al. (2010)) highlight the potentiality of NIR spectra to analyse the traceability of EVOO in order to their authentication. Casale et al. (2012) characterize PDO olive oil (qualitative information) from NIR spectra.

Moreover, this paper highlights the effect of agro-climatic conditions on spectra of olive oils. In particular, the study show the structure of linear relationships being between two sets of Big Data: NIR spectra of EVOO and agro-climatic data downloaded from the official Andalusian Automatic Weather Stations (AWEs). The graphical analysis of both data sets detects, firstly, an irregular behaviour of (original and derivative) NIR spectra corresponding to the last harvest of extraction of EVOO (H6), in particular, ten peaks of maximum discrepancy, P_1, P_2, ..., P_{10}, are determined. Secondly, the graphical analysis of the series of agro-climatic data shows irregularities in the volume of precipitation (*Precip*) or the wind speed (*WSpe*) accumulated for the previous year. This fact motivates the question about what is the effect of the agro-climatic conditions on NIR spectra or on the chemical compounds of EVOO (as NIR spectra are useful to determine quantitative information of EVOO). The answer is obtained, initially, by using bivariate analysis between the agro-climatic measurements and the spectral absorbance and, then, by extending the previous results by applying RDA. The first RDA component or redundancy axis (obtained when the matrix of spectral absorbance is the response and the matrix of agro-climatic measurements contains the explanatory variables) clearly discriminate the cases of EVOO corresponding to H6 whereas the cases corresponding of H5 are discriminated by the second RDA component. As final conclusions from bivariate and multivariate analysis, the variables monthly accumulating the precipitation (*Precip*) and the radiation (*Rad*) show, in general, the highest (in abso-

lute terms) linear correlation between the spectral absorbance, but having opposite sign. The correlation coefficients associated to wind speed (*WSpe*) are the closest to zero and so, unlike precipitations, the irregularities of the series of *WSpe* at the beginning of the harvest H6 can not be associated with the discrepancy of the EVOO NIR spectra of this harvest.

Therefore, the main contributions of this work are the treatment of the original agro-climatic data, instead of defining a factor with levels associated to the meteorological conditions, and the computational implementation in R to analyse the structure of correlations between this set of Big Data and the EVOO spectral data and efficiently represent the results (see the designed programs in the Supplementary Material). Once the effect of agro-climatic conditions on EVOO NIR spectra has been highlighted by using the Big Data and since NIR spectra contain important qualitative and quantitative information of EVOO, a further study could treat the influence of meteorological aspects in some quality parameters of olive oils, such as the fatty acids content, in order to authenticate the oils and prevent fraudulent practices.

Acknowledgements

The authors thank the financial support by 'Junta de Andalucía' (Project P08-FQM-3931).

Appendix A

Table 1: *Automatic weather stations (AWEs).*

Province	Station	Code
Cadiz	Villamartín	1
Cordova	Adamuz	2
	Baena	3
	Belmez	4
	Cabra	5
	Córdoba	6
	El Carpio	7
	Hinojosa del Duque	8
	Hornachuelos	9
	Palma del Río	10
	Santaella	11
Granada	Loja	12
	Pinos Puente	13
Jaen	Alcaudete	14
	Chiclana de Segura	15
	Jaén	16
	Higuera de Arjona	17
	Mancha Real	18
	Marmolejo	19
	Pozo Alcón	20
	San José de los Propios	21
	Santo Tomé	22
Malaga	Antequera	23
	Archidona	24
	Pizarra	25
	Sierra de Yeguas	26
Sevilla	Écija	27
	Osuna	28

Appendix B

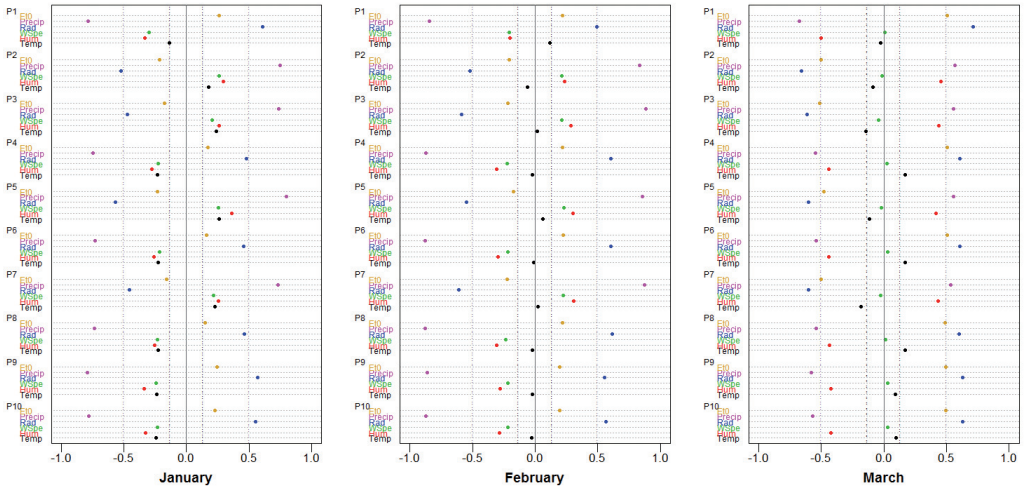


Figure B.1: *Correlations for January, February and March.*

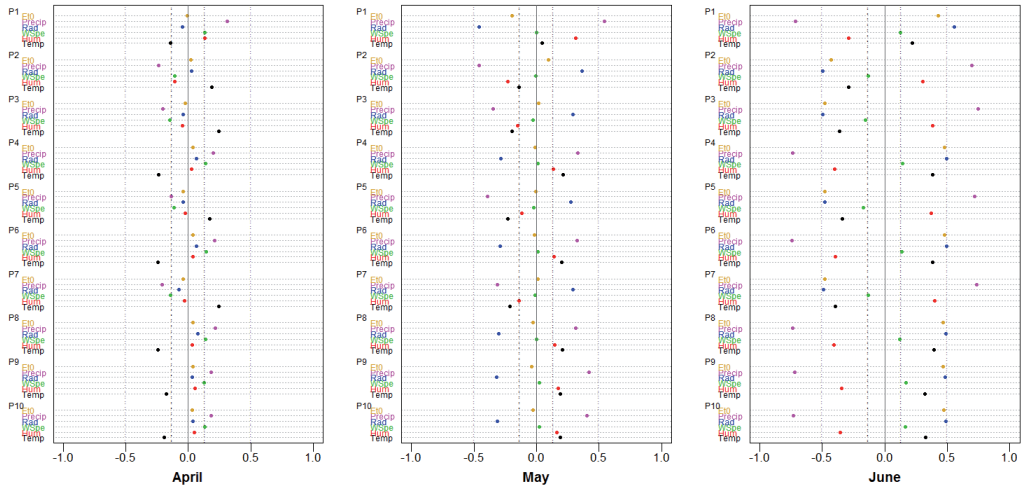


Figure B.2: *Correlations for April, May and June.*

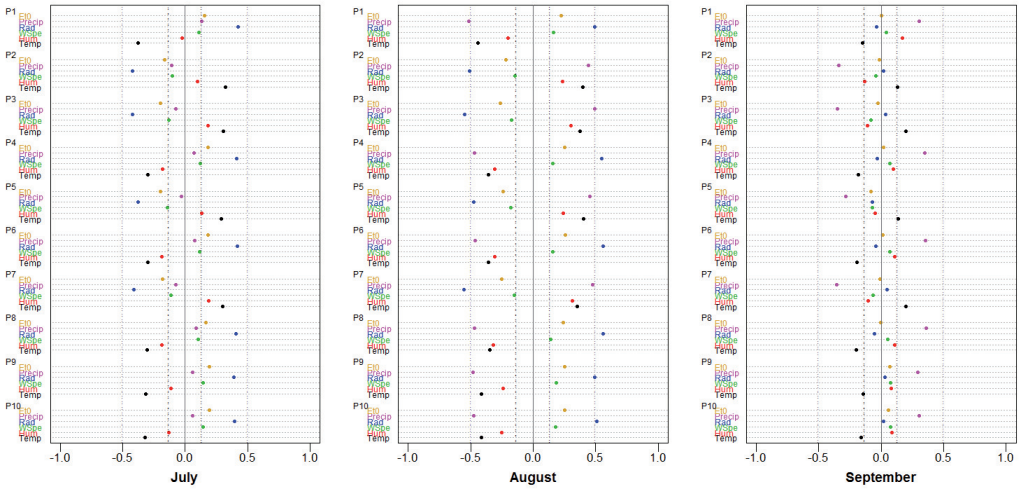


Figure B.3: *Correlations for July, August and September.*

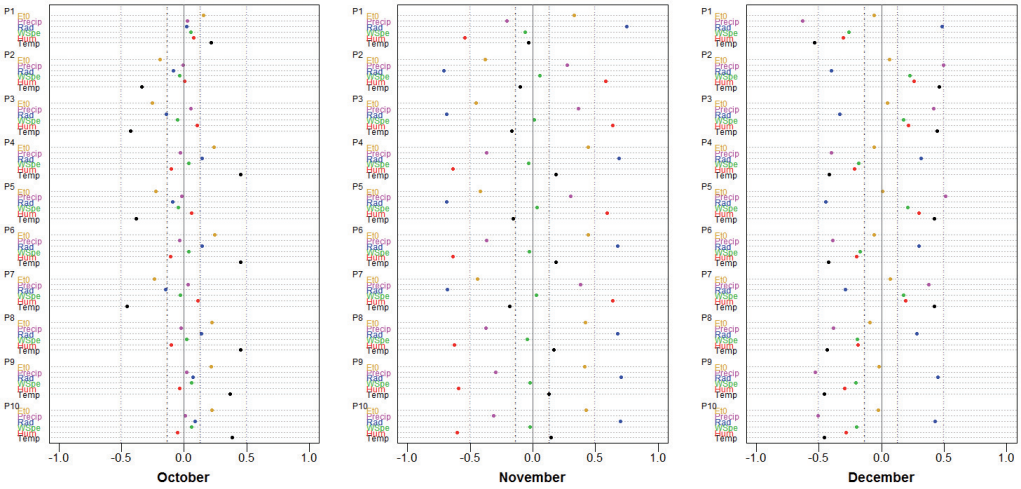


Figure B.4: *Correlations for October, November and December.*

Appendix C

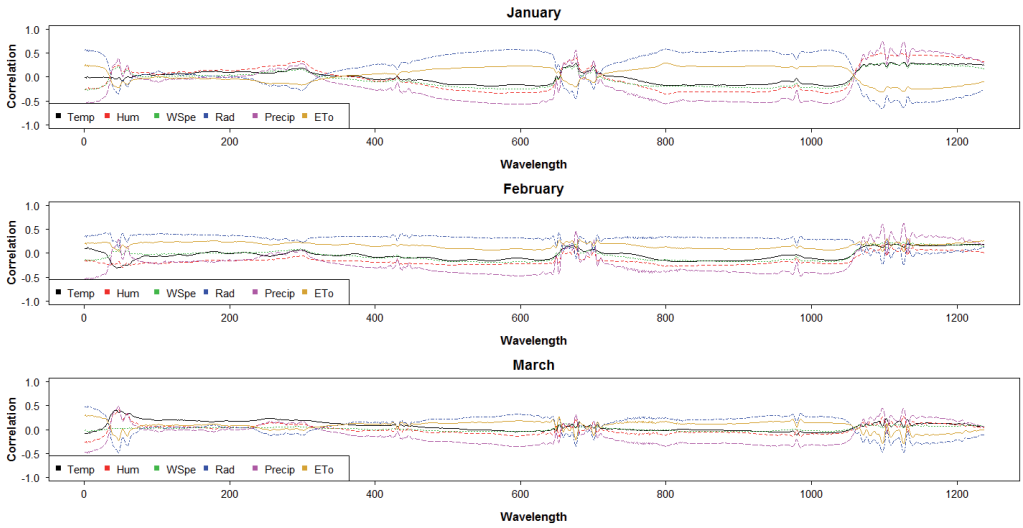


Figure C.1: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for January, February and March.*

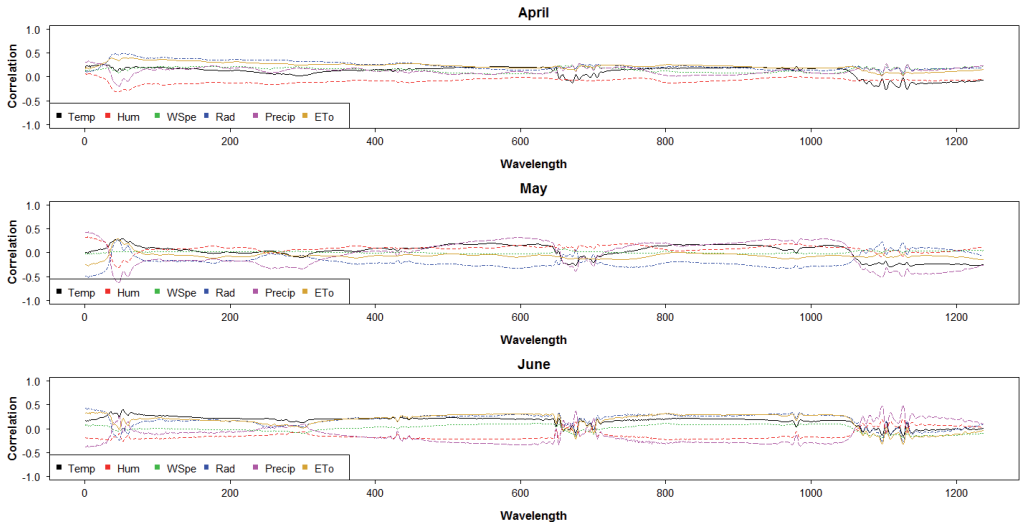


Figure C.2: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for April, May and June.*

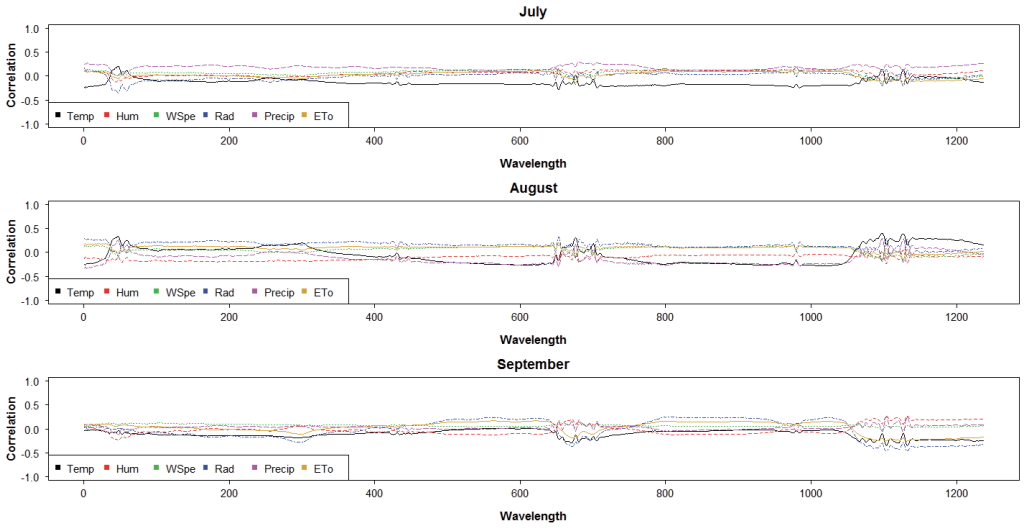


Figure C.3: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for July, August and September.*

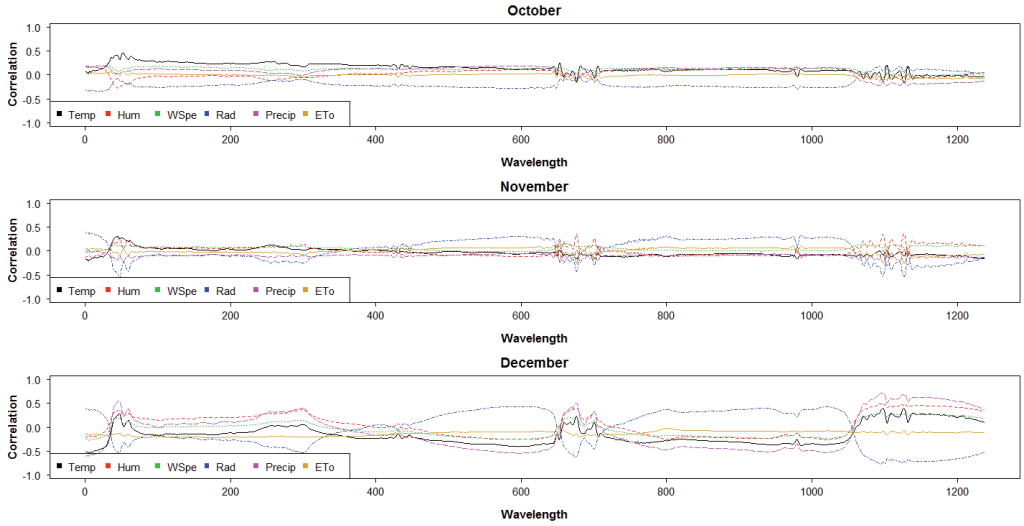


Figure C.4: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for October, November and December.*

References

Aguilera, A. M., Escabias, M., Preda, C. and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2), 289–305.

Alowaiesh, B., Singh, Z. and Kailis, S. G. (2016). Harvesting time influences fruit removal force, moisture, oil content, free fatty acids and peroxide in the oil of Frantoio and Manzanilla olive cultivars. *Australian Journal of Crop Science*, 10(12), 1662–1668. DOI:0.21475/ajcs.2016.10.12.p7737.

Awale, M., Visini, R., Probst, D., Arús-Pous, J. and Reymond, J. L. (2017). Chemical Space: Big Data Challenge for Molecular Diversity. *CHIMIA International Journal for Chemistry*, 71(10), 661–666.

Awan, A. A. (2014). Influence of agro-climatic conditions on fruit yield and oil content of olive cultivars. *Pakistan Journal of Agricultural Sciences*, 51(3).

Back, L. E. and Bretherton (2005). The relationship between wind speed and precipitation in the Pacific ITCZ. *Journal of Climate*, 18(20), 4317–4328.

Bertran, E., Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. and Montoliu, I. (2000). Near infrared spectrometry and pattern recognition as screening methods for the authentication of virgin olive oils of very close geographical origins. *Journal of Near Infrared Spectroscopy*, 8, 45.

Bradley, A. V., Gerard, F. F., Barbier, N., Weedon, G. P., Anderson, L. O., Huntingford, C., Aragão, L. E. O. C., Zelazowski, P. and Arai, E. (2011). Relationships between phenology, radiation and precipitation in the Amazon region. *Global Change Biology*, 17(6), 2245–2260.

Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., Forina, M. and Lanteri, S. (2012). Characterization of PDO olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques. *Analytical Chimica Acta*, 712, 56–63.

Chen, J. Y., Zhang, H., Ma, J., Tuchiya, T. and Miao, Y. (2015). Determination of the degree of degradation of frying rapeseed oil using Fourier-transform infrared spectroscopy combined with partial least-squares regression. *International Journal of Analytical Chemistry*, DOI: 10.1155/2015/185367.

Chiang, L., Lu, B. and Castillo, I. (2017). Big Data Analytics in Chemical Engineering. *Annual Review of Chemical and Biomolecular Engineering*, (8), 63–85.

Cornejo, V., Bueno, L. A. and Gines, I. L. (2012). Evaluation of 'Arbequina' olive oils from different growing areas of San Juan, Argentina. *VII International Symposium on Olive Growing*, 1057, 661–667.

D'Imperio, M., Mannina, L., Capitani, D., Bidet, O., Rossi, E., Bucarelli, F. M., Quaglia, G. B and Segre, A. (2007). NMR and statistical study of olive oils from Lazio: a geographical, ecological and agronomic characterization. *Food Chemistry*, 105(3), 1256–1267.

Dorey, E., Fournier, P., Léchaudel, M. and Tixier, P. (2016). Modeling sugar content of pineapple under agro-climatic conditions on Reunion Island. *European Journal of Agronomy*, 73, 64–72.

Edmunds, B. A., Clark, C. A., Villordon, A. Q. and Holmes, G. J. (2015). Relationships of preharvest weather conditions and soil factors to susceptibility of sweetpotato to postharvest decay caused by Rhizopus stolonifer and Dickeya dadantii. *Plant Disease*, 99(6), 848–857.

Falasca, S. L., Ulberich, A. C. and Ulberich, E. (2012). Developing an agro-climatic zoning model to determine potential production areas for castor bean (Ricinus communis L.).

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51(4), 1–28.

Galtier, O., Dupuy, N., Le DrÃ©au, Y., Ollivier, D., Pinatel, C., Kister, J. and Artaud, J. (2007). Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Analytica Chimica Acta*, 595(1), 136–144.

Hu, Y. and Bajorath, J. (2017). Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited. *Future Cience OA*, 3(2). DOI: 10.4155/fsoa-2017-0001.

Guo, X. (2016). Application of meteorological big data. *IEEE Communications and Information Technologies (ISCIT), 16th International Symposium*, 273–279.

Jarvis, C. K., Sapirstein, H. D., Bullock, P. R., Naeem, H. A., Angadi, S. V. and Hussain, A. (2008). Models of growing season weather impacts on breadmaking quality of spring wheat from producer fields in western Canada. *Journal of the Science of Food and Agriculture*, 88(13), 2357–2370.

Khokhar, J. S., Sareen, S., Tyagi, B. S., Singh, G., Chowdhury, A. K., Dhar, T., Sign, V., King, I. P., Young, S. D. and Broadley, M. R. (2017). Characterising variation in wheat traits under hostile soil conditions in India. *PLoS One*, 12(6), e0179208.

Legendre, P., Oksanen, J. and ter Braak, C. J. (2011). Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution*, 2(3), 269–277.

Leskinen, H. M., Suomela, J. P. and Kallio, H. P. (2009a). Effect of latitude and weather conditions on the regioisomer compositions of α-and γ-linolenoyldilinoleoylglycerol in currant seed oils. *Journal of agricultural and food chemistry*, 57(9), 3920–3926. DOI: 10.1021/jf900068b.

Leskinen, H. M., Suomela, J. P., Yang, B. and Kallio, H. P. (2009b). Regioisomer compositions of vaccenic and oleic acid containing triacylglycerols in sea buckthorn (Hippophae rhamnoides) pulp oils: influence of origin and weather conditions. *Journal of agricultural and food chemistry*, 58(1), 537–545. DOI: 10.1021/jf902679v.

Luciano, R. V., Albuquerque, J. A., Rufato, L., Miquelluti, D. J. and Warmling, M. T. (2013). Weather and soil effects on the composition of 'Cabernet Sauvignon' grape. *Pesquisa Agropecuária Brasileira*, 48(1), 97–104.

Mailer, R. J. (2004). Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. *Journal of the American Oil Chemists' Society*, 81(9), 823–827.

Martínez-Herrera, J., Siddhuraju, P., Francis, G., Davila-Ortiz, G. and Becker, K. (2006). Chemical composition, toxic/antimetabolic constituents, and effects of different treatments on their levels, in four provenances of Jatropha curcas L. from Mexico. *Food Chemistry*, 96(1), 80–89.

Merchak, N., El Bacha, E., Khouzan, R. B., Rizk, T., Akoka, S. and Bejjani, J. (2017). Geoclimatic, morphological and temporal effects on Lebanese olive oils composition and classification: A ^{1}H NMR metabolomic study, *Food Chemistry*, 217, 379–388.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs E. and Wagner, H. (2018). *vegan: Community Ecology Package. R package version 2.4-6*. https://CRAN.R-project.org/package=vegan

Orlandi, F., Bonofiglio, T., Romano, B. and Fornaciari, M. (2012). Qualitative and quantitative aspects of olive production in relation to climate in southern Italy. *Scientia Horticulturae*, 138, 151–158.

Ozdemir, Y. (2016). Effects of climate change on olive cultivation and table olive and olive oil quality. *Scientific Papers. Series B, Horticultures, LX*, 65–69.

Öztürk, B., Yalçin, A. and Özdemir D. (2010). Determination of olive oil adulteration with vegetable oils by near infrared spectroscopy coupled with multivariate calibration. *Journal of Near Infrared Spectroscopy*, 18, 191–201.

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2017). *fda: Functional Data Analysis. R package version 2.4.7*. https://CRAN.R-project.org/package=fda.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics*, Series A, 329–358.

Rymbai, H., Laxman, R. H., Dinesh, M. R., Sunoj, V. J., Ravishankar, K. V. and Jha, A. K. (2014). Diversity in leaf morphology and physiological characteristics among mango (Mangifera indica) cultivars popular in different agro-climatic regions of India. *Scientia Horticulturae*, 176, 189–193.

Sacco, A., Brescia, M. A., Liuzzi, V., Reniero, F., Guillou, G., Ghelli, S. and van der Meer, P. (2000). Characterization of Italian olive oils based on analytical and nuclear magnetic resonance determinations. *Journal of the American Oil Chemists' Society*, 77(6), 619–625.

Saeys, W., De Ketelaere, B. and Darius, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, 22(5), 335–344.

Sánchez-Rodríguez, M. I., Sánchez-López, E., Caridad, J. M., Marinas, A., Marinas, J. M. and Urbano, F. J. (2013). New insights into evaluation of regression models through a descompositon of the prediction errors: application to near-infrared spectral data. *Statistics and Operations Research Transactions Journal*, 37(1), 57–78.

Sánchez-Rodríguez, M. I., Sánchez-López, E., Marinas, A., Caridad, J. M., Urbano, F. J. and Marinas, J. M. (2014). New approaches in the chemometrics analysis of infrared spectra of extra-virgin olive oils. *Statistics and Operations Research Transactions Journal*, 38(2), 231–250.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8), 1627–1639.

Van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2), 207–219.

Veizi, A., Peçi, E. and Lazaj, L. (2016). Influence of harvesting time in chemical and organoleptic qualities of extra virgin olive oil, *Journal of Multidisciplinary Engineering Science and Technology*, 3(10), 5794–5800.

Wang, X., Song, L., Wang, G., Ren, H., Wu, T., Jia, X., Wu, H.P. and Wu, J. (2016). Operational climate prediction in the era of big data in China: Reviews and prospects. *Journal of Meteorological Research*, 30(3), 444–456.

Woodcock, T., Downey, G. and O'Donnell, C.P. (2008). Confirmation of declared provenance of European extra virgin olive oil samples by NIR spectroscopy. *Journal of Agricultural and Food Chemistry*, 56(23), 11520–11525.

Yang, W., Laaksonen, O., Kallio, H. and Yang, B. (2017). Effects of latitude and weather conditions on proanthocyanidins in berries of Finnish wild and cultivated sea buckthorn (Hippophae rhamnoides L. ssp. rhamnoides). *Food Chemistry*, 216, 87–96.

Zaied, Y. B. and Zouabi, O. (2016). Impacts of climate change on Tunisian olive oil output. *Climatic Change*, 139(3-4), 535–549.

Zheng, J., Yang, B., Ruusunen, V., Laaksonen, O., Tahvonen, R., Hellsten, J. and Kallio, H. (2012). Compositional differences of phenolic compounds between black currant (Ribes nigrum L.) cultivars and their response to latitude and weather conditions. *Journal of Agricultural and Food Chemistry*, 60(26), 6581–6593. DOI: dx.doi.org/10.1021/jf3012739.

SORT 42 (2) July-December 2018, 237-252 DOI: 10.2436/20.8080.02.76

Poisson excess relative risk models: new implementations and software

Manuel Higueras[1,2,3] and Adam Howes[2]

Abstract

Two new implementations for fitting Poisson excess relative risk methods are proposed for assumed simple models. This allows for estimation of the excess relative risk associated with a unique exposure, where the background risk is modelled by a unique categorical variable, for example gender or attained age levels. Additionally, it is shown how to fit general Poisson linear relative risk models in R. Both simple methods and the R fitting are illustrated in three examples. The first two examples are from the radiation epidemiology literature. Data in the third example are randomly generated with the purpose of sharing it jointly with the R scripts.

MSC: 62J02

Keywords: Radiation epidemiology, Poisson non-linear regression, improper priors, R programming

1. Introduction

The excess relative risk (ERR) represents the additional risk of disease (e.g., leukaemia, brain tumour) per unit of exposure (e.g., absorbed dose of ionising radiation). In a linear ERR model with d exposures, the risk is modelled by

$$e^{\eta}\left(1+\sum_{j=1}^{d}\beta_j D^{(j)}\right),$$

where each parameter β_j is the ERR associated with the absorbed dose $D^{(j)}$. The risk is represented by the product of the background risk term, e^{η}, and the term within parenthesis, which is the relative risk. Poisson linear ERR models can be used to calculate the ERR in longitudinal cohort studies with active follow-up. It is assumed that

[1] Departamento de Matemáticas y Computación, Universidad de La Rioja, Edificio CCT - C/ Madre de Dios 53, 26006 Logroño (La Rioja), Spain.

[2] Basque Center for Applied Mathematics, Bilbao, Spain.

[3] Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK.

Received: January 2018
Accepted: November 2018

$$C_i \sim \text{Pois}\left(PY_i e^{\eta_i}\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right), \tag{1}$$

where C_i and PY_i are the number of disease cases and the number of person-years of follow-up, and $D_i^{(j)}$ is the mean dose (weighted by the person-years) of exposure j for stratum $i = \{1,\ldots,n\}$ respectively (BEIR VII Phase 2, 2006). The most common situation in ERR models is to have only one exposure variable. More complicated ERR models with effect modification of the dose-response are also often reported, e.g. Grant et al. (2017).

The background risk can be modelled by m covariates, *i.e.* $\eta_i = \alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)}$. These covariates are usually time-dependent variables, e.g. attained age or transplant status. The model (1) is not the canonical log-linear Poisson model (McCullagh and Nelder, 1989). Since it mixes both log-linear and linear terms it is a generalised non-linear model.

In this work, Poisson ERR models with simple forms are studied to obtain estimates in closed or almost closed form. This allows calculations to be made faster and more accurate. As an alternative to other implementations in the literature, such as Epicure (Preston et al., 1993) and SAS (SAS Institute Inc., Cary, North Carolina) (Richardson, 2008), the software R (R Core Team, 2017) was used to fit general Poisson ERR models. Three applied examples are detailed, and the data and R scripts of the third example are included as supplementary material.

2. Simple ERR model

A simple ERR model may be defined by assuming one exposure, $d = 1$, and that the background risk linear predictor, η_i, is of the form $\alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)$, where x_i represents a categorical variable with K levels. This model is simple, with only one exposure, and one categorical covariate in the background risk term.

Following these assumptions

$$C_i \sim \text{Pois}\left(PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}\left(1+\beta D_i\right)\right). \tag{2}$$

Let $\vec{\alpha} = \{\alpha_1,\ldots,\alpha_K\}$ and $X = \{C, PY, D, x\}$, then the likelihood of the parameter set $\Theta = \{\vec{\alpha}\beta\}$ is given by

$$\mathrm{L}(\Theta|X) = \prod_{i=1}^{n}\frac{[PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)]^{C_i}\exp(-PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))}{C_i!} \tag{3}$$

and the log-likelihood is

$$
\begin{aligned}
l(\Theta|X) = \quad & \log(\mathrm{L}(\Theta|X)) = \sum_{i=1}^{n}\left[C_i(\log PY_i + \alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i) + \log(1+\beta D_i))\right] \\
- \quad & \sum_{i=1}^{n} PY_i \mathrm{e}^{\alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i) - \sum_{i=1}^{n}\log C_i!.
\end{aligned}
\tag{4}
$$

For this model two implementations are proposed, one is frequentist and the other is Bayesian. The frequentist implementation provides a closed form of the profile likelihood of β and the Bayesian provides the marginal posterior for β also in a closed form.

2.1. Profile likelihood and maximum likelihood estimator

A profile likelihood CI (PLCI) for the ERR parameter is preferred to the typical Wald CI because the likelihood function of ERR models is usually non-normal in shape. Let $\mathrm{L}(\Theta|X)$ be the likelihood function as in 3, then the profile likelihood of β is

$$L_1(\beta|X) = \max_{\theta} \mathrm{L}(\theta, \beta|X).$$

The $(1-a)\cdot 100\%$ PLCI are the values of β that meet the requirement

$$\log(L_1(\beta|X)) > \hat{l} - \chi^2_{1,1-a}/2,$$

where $\hat{l} = l(\hat{\Theta}|X)$ is the maximum value of the log-likelihood function and $\chi^2_{1,1-a}$ is the $1-a$ quantile of a chi-squared distribution with 1 degree of freedom. Note that $L_1(\beta|X)$ is the likelihood of a Poisson GLM: $C \sim \mathrm{Pois}\,(PY(1+\beta D)\mathrm{e}^{\eta})$ where $PY(1+\beta D)$ is the offset. In general, for only one exposure the profile likelihood of the ERR is the likelihood of a Poisson GLM with canonical logarithm link.

Assuming a simple model as (2), the profile likelihood for β can be calculated by solving

$$
\begin{cases}
\dfrac{\partial l}{\partial \alpha_1} = & S - \mathrm{e}^{\alpha_1}\left(T_1 + \displaystyle\sum_{k=2}^{K} T_k \mathrm{e}^{\alpha_k}\right) = 0 \\
\dfrac{\partial l}{\partial \alpha_k} = & S_k - T_k \mathrm{e}^{\alpha_1+\alpha_k} = 0,\ k = \{2,\dots,K\}
\end{cases},
$$

where $S = \sum_{i=1}^{n} C_i$, $S_k = \sum_{i|x_i=k} C_i$, and $T_k = \sum_{i|x_i=k} PY_i(1+\beta D_i)$. Let $\vec{\alpha}(\beta) = \{\alpha_1(\beta), \dots, \alpha_K(\beta)\}$ then the profile likelihood for β is $L_1(\beta) = L(\vec{\alpha}(\beta), \beta|X)$ where

$$
\begin{aligned}
\alpha_1(\beta) = \quad & \log\left(S - \sum_{k=2}^{K} S_k\right) - \log(T_1), \\
\alpha_k(\beta) = \quad & \log(S_k) - \log(T_k) - \alpha_1(\beta),\ k = \{2,\dots,K\}.
\end{aligned}
$$

To obtain the maximum likelihood estimators of the parameters, the partial derivative of the log-likelihood with respect β is evaluated at $\vec{\alpha} = \vec{\alpha}(\beta)$, *i.e.*

$$\left.\frac{\partial l}{\partial \beta}\right|_{\vec{\alpha}=\vec{\alpha}(\beta)} = \sum_{i=1}^{n} \frac{C_i D_i}{1+\beta D_i} - \mathrm{e}^{\alpha_1(\beta)} R_1 - \mathrm{e}^{\alpha_1(\beta)} \sum_{k=2}^{K} \mathrm{e}^{\alpha_k(\beta)} R_k = 0,$$

where $R_k = \sum_{i|x_i=k} PY_i D_i$. This equation is solved numerically to get the estimator $\hat{\beta}$ and the rest of the estimators are $\hat{\vec{\alpha}} = \vec{\alpha}(\hat{\beta})$.

The likelihood ratio test p-value for null hypothesis $\beta = 0$ is

$$\mathrm{P}\left(\chi_1^2 > l(\vec{\alpha}(\hat{\beta}), \hat{\beta}|X) - l\left(\vec{\alpha}(0), 0|X\right)\right),$$

where χ_1^2 is a chi-squared distribution with 1 degree of freedom.

It is possible that the PLCI bound does not converge. In this situation, the Wald-type CI bound is usually reported. This can be calculated by the Hessian matrix,

$$H(\vec{\alpha}, \beta) =$$
$$\begin{bmatrix} -\mathrm{e}^{\alpha_1}\left(T_1 + \sum_{k=2}^{K} \mathrm{e}^{\alpha_k} T_k\right) & -\mathrm{e}^{\alpha_1+\alpha_2} T_2 & -\mathrm{e}^{\alpha_1+\alpha_3} T_3 & \dots & -\mathrm{e}^{\alpha_1+\alpha_K} T_K & -\mathrm{e}^{\alpha_1}\left(R_1 + \sum_{k=2}^{K} \mathrm{e}^{\alpha_k} R_k\right) \\ -\mathrm{e}^{\alpha_1+\alpha_2} T_2 & 0 & 0 & \dots & 0 & -\mathrm{e}^{\alpha_1+\alpha_2} R_2 \\ -\mathrm{e}^{\alpha_1+\alpha_3} T_3 & 0 & 0 & \dots & 0 & -\mathrm{e}^{\alpha_1+\alpha_3} R_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\mathrm{e}^{\alpha_1+\alpha_K} T_K & -\mathrm{e}^{\alpha_1+\alpha_K} R_K & & & & \\ -\sum_{i=1}^{n} \frac{C_i D_i^2}{(1+\beta D_i)^2} & & & & & \end{bmatrix},$$

and evaluating it at the maximum likelihood estimator, *i.e.* $H(\vec{\alpha}(\hat{\beta}), \hat{\beta})$. The variance-covariance matrix is $-H(\vec{\alpha}(\hat{\beta}), \hat{\beta})^{-1}$.

2.2. Posterior ERR

Bayesian analysis combines prior information, in the form of probability distributions, with the likelihood function of an assumed model, providing posterior results as probability distributions too. The continuous version of Bayes' theorem establishes

$$\mathrm{P}(\Theta|X) = \frac{\mathrm{L}(\Theta|X)\mathrm{P}(\Theta)}{\int \mathrm{L}(\Theta|X)\mathrm{P}(\Theta)\mathrm{d}\Theta}, \tag{5}$$

where Θ is the continuous parameter set, X is the observed data set, $\mathrm{L}(\Theta|X)$ is the likelihood function, $\mathrm{P}(\Theta)$ is the prior probability density function of Θ and $\mathrm{P}(\Theta|X)$ is the posterior probability density of Θ given data X. See, for instance Christensen et al. (2011), for further description.

Following model (2): X, Θ and $\mathrm{L}(\Theta|X)$ are as stated in Section 2. Assuming all α_k's and β are independent, the prior probability density is

$$\mathrm{P}(\Theta) = \mathrm{P}(\beta)\prod_{k=1}^{K}\mathrm{P}(\alpha_k).$$

It is also assumed that all α_k's priors are non informative, such that the probability is the same for all the values in the support of the parameters. This leads to the following improper uniform priors:

$$\alpha_k \quad \sim \mathscr{U}(-\infty,+\infty), k = \{1\dots K\} \tag{6}$$

and a prior for β open to any distribution with support bounded below by $-1/\max(D)$, to ensure the Poisson intensity is positive. The Bayesian framework affords the definition of improper prior distributions.

Applying Bayes' theorem (5), the posterior of Θ is

$$\begin{aligned}
&\mathrm{P}(\Theta|X) \propto \mathrm{P}(\beta)\cdot\mathrm{L}(\Theta|X)\\
&\propto \mathrm{P}(\beta)\prod_{i=1}^{n}(PY_i\mathrm{e}^{\alpha_1+\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}\exp(-PY_i\mathrm{e}^{\alpha_1+\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))\\
&= \mathrm{P}(\beta)\cdot\exp\left(S\alpha_1-\mathrm{e}^{\alpha_1}\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right)\cdot\\
&\cdot\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}.
\end{aligned} \tag{7}$$

The goal here is to get the marginal posterior of the ERR, the posterior distribution of β. Let $\vec{\alpha}_{-1} = (\alpha_2,\dots,\alpha_K)$, the first step is to calculate the joint marginal posterior of $(\vec{\alpha}_{-1},\beta)$ which it is proportional to the integral of expression (7) over α_1, *i.e.*

$$\begin{aligned}
\mathrm{P}(\vec{\alpha}_{-1},\beta|X) \propto\ & \mathrm{P}(\beta)\int_{-\infty}^{+\infty}\mathrm{L}(\Theta|X)\mathrm{d}\alpha_1\\
=\ & \mathrm{P}(\beta)\left[\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}\right]\\
\cdot\ & \int_{-\infty}^{+\infty}\exp\left(S\alpha_1-\mathrm{e}^{\alpha_1}\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right)\mathrm{d}\alpha_1
\end{aligned}$$

$$\begin{aligned}
&= \mathrm{P}(\beta)\frac{\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}}{\left[\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right]^{S}}(S-1)! \\
&\propto \frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\mathrm{e}^{\sum_{k=2}^{K}S_k\alpha_k}}{\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)+\sum_{k=2}^{K}\left(\mathrm{e}^{\alpha_k}\sum_{i|x_i=k}PY_i(1+\beta D_i)\right)\right]^{S}}.
\end{aligned} \tag{8}$$

Then the marginal posterior of the ERR is proportional to the multiple integral of Expression (8) over $\vec{\alpha}_{-1}$,

$$\begin{aligned}
\mathrm{P}(\beta|X) = &\int_{\vec{\alpha}_{-1}}\mathrm{P}(\vec{\alpha}_{-1},\beta|X)\mathrm{d}\vec{\alpha}_{-1} \\
\propto\ &\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\int_{\vec{\alpha}_{-1}}\mathrm{e}^{\sum_{k=2}^{K}S_k\alpha_k} \\
&\cdot\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)+\sum_{k=2}^{K}\left(\mathrm{e}^{\alpha_k}\sum_{i|x_i=k}PY_i(1+\beta D_i)\right)\right]^{-S}\mathrm{d}\vec{\alpha}_{-1} \\
= &\frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K}S_k-S}}{\prod_{k=2}^{K}\left[\sum_{i|x_i=k}PY_i(1+\beta D_i)\right]^{S_k}} \\
&\cdot\frac{(S_2-1)!}{\prod_{i=1}^{S_2-1}S-i}\prod_{k=3}^{K}\frac{(S_k-1)!}{\prod_{i=1}^{S_k-1}S-\sum_{j=2}^{k-1}S_k-i} \\
\propto\ &\frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K}S_k-S}}{\prod_{k=2}^{K}\left[\sum_{i|x_i=k}PY_i(1+\beta D_i)\right]^{S_k}}.
\end{aligned} \tag{9}$$

Consequently,

$$\mathrm{P}(\beta|X) = \frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1} PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K} S_k - S}}{N\prod_{k=2}^{K}\left[\sum_{i|x_i=k} PY_i(1+\beta D_i)\right]^{S_k}}, \tag{10}$$

where N is the normalizing constant

$$N = \int_0^{+\infty} \mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1} PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K} S_k - S} \prod_{k=2}^{K}\left[\sum_{i|x_i=k} PY_i(1+\beta D_i)\right]^{-S_k} \mathrm{d}\beta, \tag{11}$$

that is calculated by numerical integration (there is no analytical solution). The probability density (10) does not have a recognizable form, but this is not unusual when dealing with Bayesian analysis.

The integrals in expressions (8) and (9) are calculated by recursive integration by parts.

3. Poisson ERR fitting in R

Cohort studies in radiation epidemiology are usually huge, and hence maximum likelihood estimation of the model parameters is computationally intensive. This computational load increases for the calculation of the profile likelihood confidence intervals.

As mentioned in Section 1 a general ERR model has the form

$$C_i \sim \mathrm{Pois}\left(PY_i \mathrm{e}^{\alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)}}\left(1 + \sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right). \tag{12}$$

Let $\vec{\alpha} = \{\alpha_0, \ldots, \alpha_m\}$ and $\vec{\beta} = \{\beta_1, \ldots, \beta_d\}$, the log-likelihood function of of the parameter set $\Theta = \left\{\vec{\alpha}, \vec{\beta}\right\}$

$$
\begin{aligned}
l(\Theta|X) = & \sum_{i=1}^{n}\left[C_i(\log PY_i + \alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)} + \log\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right] \\
& - \sum_{i=1}^{n} PY_i \mathrm{e}^{\alpha_0+\sum_{k=1}^{m}\alpha_k x_i^{(k)}}\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right) - \sum_{i=1}^{n}\log C_i!.
\end{aligned} \tag{13}
$$

The gradient of the log-likelihood function can be efficiently defined by the following expressions

$$
\begin{aligned}
\frac{\partial l}{\partial \vec{\alpha}} &= \vec{S} - [(PY \circ (1+\vec{\beta}\cdot\mathscr{D}) \circ E)\cdot A], \\
\frac{\partial l}{\partial \vec{\beta}} &= [C \oslash (1+\vec{\beta}\cdot\mathscr{D})]\cdot\mathscr{D} - (PY\circ E)\cdot\mathscr{D},
\end{aligned} \tag{14}
$$

where

$$
\mathscr{D} = \begin{bmatrix} D_1^{(1)} & D_1^{(2)} & \dots & D_1^{(d)} \\ D_2^{(1)} & D_2^{(2)} & \dots & D_2^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ D_n^{(1)} & D_n^{(2)} & \dots & D_n^{(d)} \end{bmatrix}, \; A = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{bmatrix},
$$

$$
E = \exp\left(\vec{\alpha}\cdot A^{\mathrm{T}}\right), \; \vec{S} = C\cdot A,
$$

and operators $\circ$ and $\oslash$ represents the Hadamard product and division respectively.

In cases where the PLCI bound does not converge, the Hessian can be calculated using the following second-order partial derivatives of the log-likelihood to calculate the Wald-type CI bound,

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\alpha_t\partial\alpha_q} &= -\mathrm{e}^{\alpha_0} T_{t,q},\; t,q = \{0,\dots,m\}, \\
\frac{\partial^2 l}{\partial\alpha_t\partial\beta_q} &= -\mathrm{e}^{\alpha_0} R_{t,q},\; t = \{0,\dots,m\},\; q = \{1,\dots,d\}, \\
\frac{\partial^2 l}{\partial\beta_t\partial\beta_q} &= -\sum_{i=1}^{n}\frac{C_i D_i^{(t)} D_i^{(q)}}{\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)^2},\; t,q = \{1,\dots,d\},
\end{aligned}
$$

where

$$T_{t,q} = \sum_{i=1}^{n} PY_i \left(1 + \sum_{j=1}^{d} \beta_j D_i^{(j)}\right) x_i^{(t)} x_i^{(q)} \mathrm{e}^{\sum_{k=1}^{m} \alpha_k x_i^{(k)}},$$

$$R_{t,q} = \sum_{i=1}^{n} PY_i D_i^{(q)} x_i^{(t)} \mathrm{e}^{\sum_{k=1}^{m} \alpha_k x_i^{(k)}},$$

In R (version 3.5.1), by means of the `maxLik()` function from the `maxLik` package (Henningsen and Toomet, 2011) (version 1.3.4), model (12) can be fitted by defining the log-likelihood function (13). For faster and more accurate results, the gradient function implemented as in (14) can be included in the `maxLik()` function.

The R script for the results in Section 4.3 is provided as supplementary material, as a reference for the R implementation of model (12) fitting.

ERR models are usually fitted by Epicure, a very specialised proprietary software, which is the gold standard in radiation epidemiology practice. In recent years, some studies have been published using SAS, e.g. Journy et al. (2015), but there is not a SAS Stored Process for this aim. However, there are some SAS macros for fitting ERR models and calculating PLCI's, e.g. in Richardson (2008) for Poisson models by means of `PROC NLMIXED`. In Grant et al. (2017), an R routine was developed to analyze the Life Span Study data of a-bomb survivors in Hiroshima and Nagasaki, by means of the `gnm()` function in package `gnm` (Turner and Firth, 2018). There is also an R package called `linERR` which fits ERR models for censored survival data (Morinña, 2016).

This is proposed as a free licence and open source alternative of Epicure's `AMFIT` module, which is used to fit Poisson ERR models. Moreover, the R routines in Grant et al. (2017) also cover this purpose, in fact they also allow to fit more complex models with dose-effect modification.

The previous step to fitting the Poisson ERR model is to generate the person-years table. These tables are created by stratifying by categories of different variables, e.g. attained age, the original censored data. For each cell of the table, the accumulated person-years and events are calculated. In Epicure the module `DATAB` generates these tables. Further work in this project includes the creation of an R package with tools to fit Poisson ERR models, calculate PLCI's and generate person-years tables. Function `pyears()` in the `survival` package (Therneau, 2015) builds person-time tables, but for non-dynamic exposures.

4. Practical examples

Two applied examples for data from the literature are given. The third example is the application of the proposed implementations here to a subset from the first example data set. This third example is presented to facilitate reproducible and replicable research, because the data sets of the first two examples are not shareable.

The ERR is per mGy (milligray) in all examples shown here.

4.1. Pearce et al. 2012

Pearce et al. (2012) analysed the risk of leukaemia and brain tumours in young patients who were first underwent computed tomography (CT) scans in National Health Service hospitals in England, Wales, or Scotland in a 23 years retrospective cohort study. In the leukaemia follow-up, there were 74 leukaemia diagnosis for 178,604 patients, and a total of 1,720,984 person-years. The person-year table was built assuming 2 years exclusion and lag periods.

A Poisson ERR model is assumed with unique exposure (the accumulated ionising radiation dose), and the background risk is modelled by

$$\eta = \alpha_1 \mathbf{1}_{a_i<5} + \alpha_2 \mathbf{1}_{5\leq a_i<20} + \alpha_3 \mathbf{1}_{20\leq a_i<30} + \alpha_4 \mathbf{1}_{30\leq a_i<35} + \alpha_5 \mathbf{1}_{a_i\geq 35}$$

where a_i is the attained age. This model has the same form as the simple model in Section 2: one exposure and baseline rate modelled by a unique categorical variable.

Following the implementation in Section 2.1, the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0362$ and its 95% PLCI is $(0.0052, 0.1198)$ with p-value 0.0097. These values match with those shown in Pearce et al. (2012).

Following the implementation in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0015, +\infty)$, Figure 1 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0361 and its 95% highest posterior density (HPD) interval is (0.0023, 0.1460).

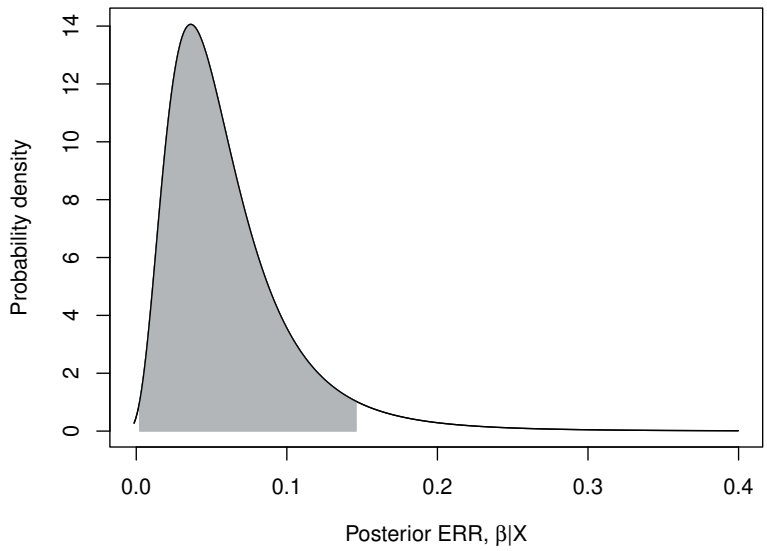


Figure 1: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.1.*

One of the big advantages of the Bayesian framework is that it is possible to calculate the posterior probability of a parameter being contained by a given interval. For instance in this case there is a posterior probability of 0.5111 for the ERR being greater than 0.050.

In order to compare this model with improper flat priors to a model with informative priors, a model with the following priors is assumed

$$\begin{aligned} \vec{\alpha} &\sim \mathcal{N}([-10,0,0,0,0]^{\mathrm{T}}, 0.1 \cdot I_5), \\ \beta &\sim \text{Gamma}(1.1, 5). \end{aligned} \tag{15}$$

I_5 is the identity matrix of size 5. The parametrization of the multivariate normal distribution is represented by the mean vector and precision matrix, and for the gamma distribution by the shape and rate values. The posterior distribution of the ERR is drawn using JAGS (version 4.3.0) (Plummer (2003)). The modal posterior ERR value is 0.0377 and its 95% HPD interval is (0.0042, 0.114). This MCMC model has 2 chains of 50,000 iterations after 1000 burning iterations and thinning interval 10. It is computational intensive, it takes around 20 hours.

Although both the Bayesian and the frequentist methods provide estimation and uncertainty results, when comparing them it is important to note that they represent different foundational approaches. In particular, the frequentist method assumes that the parameter is a fixed value and the maximum likelihood estimator is a random variable whereas the Bayesian method assumes the opposite.

4.2. Harbron et al. 2018

Harbron et al. (2018) analysed the risk of leukaemia and lymphoma in patients who underwent cardiac catheterizations while aged 22 years or younger. There were 36 cases for 9,467 patients, and a total of 74,405.88 person-years at risk in this study. Doses were lagged by 2 years. The exclusion period was also 2 years.

To calculate the ERR, a Poisson ERR model was assumed with unique exposure with background risk as

$$\eta = \alpha_1 \mathbf{1}_{a_i<5} + \alpha_2 \mathbf{1}_{5\leq a_i<10} + \alpha_3 \mathbf{1}_{10\leq a_i<15} + \alpha_4 \mathbf{1}_{15\leq a_i<20} + \alpha_5 \mathbf{1}_{20\leq a_i<25} + \alpha_6 \mathbf{1}_{a_i\geq 25} + T_i$$

where a_i is the attained age and T_i represents the status of organ transplantation. Note that this model does not have the same structure as the simple model in Section 2.

In Harbron et al. (2018) this model was fitted in R as stated in Section 3. The maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0180$, and its 95% PLCI is $(-0.0021, 0.0961)$ with p-value 0.1084.

Assuming a simple model with background risk modelled only by the transplant status, the results for the two methods are:

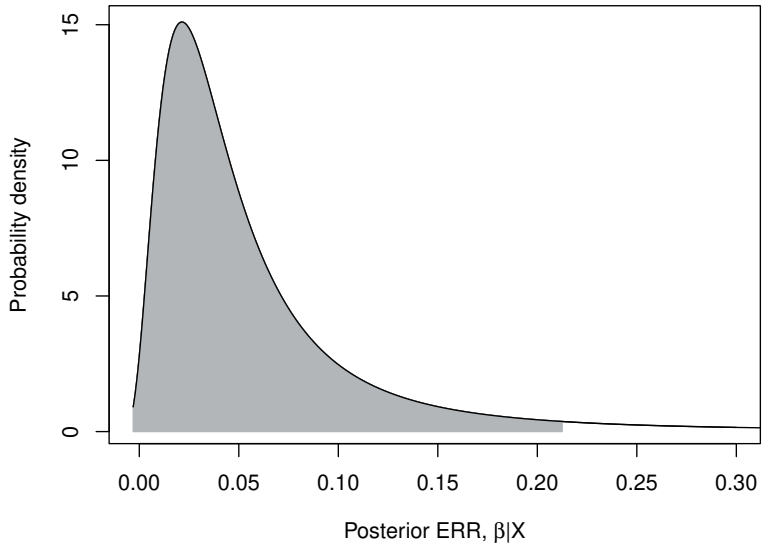


Figure 2: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.2.*

- Following the method in Section 2.1 the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0214$ and its 95% PLCI is $(-0.0008, 0.1049)$ with p-value = 0.0661.
- Following the method in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0030, 1)$, Figure 2 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0215 and its 95% HPD is $(-0.0030, 0.2125)$, and $\mathrm{P}(\beta|X > 0.050) = 0.4091$.

4.3. Sub-cohort

A 14,000 random row subset of the person-years table from the leukaemia analysis in Section 4.1, with information of accumulated person-years, weighted mean accumulated dose, sex and weighted mean attained age was generated. In this sub-cohort there are 9 leukaemia cases in a total of 158,953.3 person-years.

A Poisson ERR model is assumed, with unique exposure and background risk modelled by $\eta = \alpha_0 + \alpha_1 a_i$, where a_i is the attained age. The attained age is not a categorical variable, so this model does not have the same structure as the simple model in Section 2.

Fitting this model in R as stated in Section 3, the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0247$ (it agrees with the result returned by `gnm()`), and its 95% PLCI is $(-0.0553^*, 0.3341)$ with p-value 0.4535. The symbol * denotes the bound is Wald-type.

To check the effect of gender on the ERR, an interaction between the dose and the sex is added to the previous model, *i.e.* the ERR term is $(1 + \beta_1 D_i + \beta_2 F_i D_i$ where F_i is the indicator of female patient.

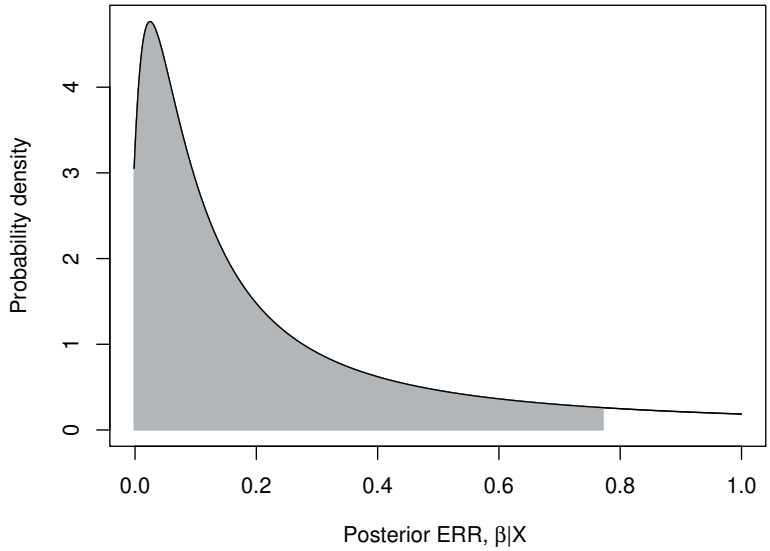


Figure 3: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.3.*

This can be fitted as a model with two exposures, D_i and F_iD_i, and the ERR results 0.0039 for male and 0.0541 for female, this is $\hat{\beta}_1 = 0.0039$ and $\hat{\beta}_2 = 0.0502$, but the female effect is not significant because the likelihood ratio test p-value for testing $\beta_2 = 0$ is 0.3059.

Now, assuming a simple model with background risk modelled by three categories of attained age, *i.e.*

$$\eta = \alpha_1 \mathbf{1}_{a_i<10} + \alpha_2 \mathbf{1}_{10\leq a_i<15} + \alpha_3 \mathbf{1}_{a_i\geq 15},$$

the results for the two methods are:

- Following the method in Section 2.1 the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0247$ and its 95% PLCI is (-0.0584^*, 0.3659) with p-value = 0.3884.
- Following the method in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0015, 1)$, Figure 3 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0247 and its 95% HPD interval is (-0.0015, 0.7717), and $\mathrm{P}(\beta|X > 0.050) = 0.7724$. If $\beta \sim \mathrm{Gamma}(1.1, 5)$, the modal posterior ERR value is 0.0234 and its 95% HPD interval is (0, 0.3294). Analogously to example at Section 4.1, an MCMC model is applied to draw the posterior of the ERR, assuming the same priors (with the difference of the dimension of $\vec{\alpha}$, *i.e.* $\vec{\alpha} \sim \mathscr{N}([-10,0,0,0,0]^{\mathrm{T}}, 0.1 \cdot I_3)$), the modal posterior ERR value is 0.0346 and its 95% HPD interval is (0.0001, 0.3073).

5. Conclusion

The simple methods presented here for estimating the ERR in radiation epidemiology follow-up studies are easy to implement. Although these models have restricted forms, they cover a wide range of situations. For instance, the leukaemia analysis in Pearce et al. (2012) was performed with this type of model. Additionally, they can be used to get sensible initial values for fitting ERR models with more complex structures.

R is an open-source statistical software program with a free license and large user community. As such, it is well suited for the development of reproducible and replicable research. In this work an R script for fitting Poisson ERR models is shared and guidelines for implementing ERR models in R are given in Section 3.

Further work in this project will lead to the development of an R package with tools to fit Poisson ERR models, build person-years tables with time-dependent variables, and calculate PLCI's. This package will have application in radiation epidemiology follow-up studies.

Acknowledgments

This research was supported by the Basque Government through the BERC 360 2014-2017 and the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and MINECO Challenges MTM2017-82379-R.

MH is grateful for the help of Dr. Richard W. Harbron and Prof. Mark S. Pearce, from Newcastle University; and to the reviewers for their valuable reviews.

References

Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation (2006). *Health Risks from exposure to low levels of ionizing radiation. BEIR VII Phase 2*. Washington: The National Academies Press.

Christensen, R., Johnson, W., Brasncum, A. and Hanson, T.E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton: Champman & Hall/CRC Press.

Grant, E.J., Brenner, A., Sugiyama, H., Sakata, R., Sadakane, A., Utada, M., Cahoon, E. K., Milder, C. M., Soda, M., Cullings, H. M., Preston, D. L., Mabuchi, K. and Ozasa, K. (2017). Solid Cancer Incidence among the Life Span Study of Atomic Bomb Survivors: 1958–2009. *Radiation Research*, 187(5), 513–537.

Harbron, R.W., Chapple, C.-L., O'Sullivan, J.J., Lee, C, McHugh, K., Higueras, M. and Pearce, M.S. (2018). Cancer incidence among children and young adults who have undergone x-ray guided cardiac catheterization procedures. *European Journal of Epidemiology*, 33(4), 393–401.

Henningsen, A. and Toomet, O. (2011). `maxLik`: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), 443–458.

Journy, N., Rehel, J.-L., Ducou Le Pointe, H., Lee, C., Brisse, H., Chateil, J.-F., Caer-Lorho, S., Laurier, D. and Bernier, M.-O. (2015). Are the studies on cancer risk from CT scans biased by indication? Elements of answer from a large-scale cohort study in France. *British Journal of Cancer*, 112(1), 185–193.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. Boca Raton: Champman & Hall/CRC Press.

Morinña, D. (2016). *linERR: Linear Excess Relative Risk Model*, version 1.0, URL: https://CRAN.R-project.org/package=linERR.

Pearce, M.S., Salotti, J.A., Little, M.P., Mchugh, K., Lee, C., Kim, K.P., Howe, N.L., Ronckers, C.M., Rajaraman, P., Craft, A.W., Parker, L. and Berrington de González, A. (2012). Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet*, 380(9840), 499–505.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

Preston, D.L., Lubin, J.H., Pierce, D.A. and McConney, M.E. (1993). *Epicure: user's guide*. Seattle: Hirosoft International Corporation.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Richardson, D.B. (2008). A simple approach for fitting linear relative rate models in SAS. *American Journal of Epidemiology*, 168(11), 1333–1338.

Therneau, T. (2015). *A Package for Survival Analysis in S*, version 2.38, URL: https://CRAN.R-project.org/package=survival.

Turner, H. and Firth, D. (2018). *Generalized nonlinear models in R: An overview of the gnm package*, version 1.1-0, URL: https://cran.r-project.org/package=gnm.

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

SORT Statistics and Operations Research Transactions
Institut d'Estadstica de Catalunya (Idescat)
Via Laietana, 58 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT *(Statistics and Operations Research Transactions)*

Name _____

Organisation _____

Street Address _____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2018 (volume 42)

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:
sort@idescat.cat

Aims

SORT (*Statistics and Operations Research Transactions*) -formerly *Qüestiió*- is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society and the Catalan Statistical Society. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, Índice *Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada* e *Investigación Operativa*, published by the UPC till 1977.

Institut d'Estadística de Catalunya

The Statistical Institute of Catalonia (Idescat) is the statistical office of the Government of Catalonia.
Its main duty is the management of the Catalan Statistical System by planning, coordinating and standardizing the statistical activity as well as providing statistical technical assistance.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
www.idescat.cat/sort/