Statistical Science
2024, Vol. 39, No. 2, 262–285
https://doi.org/10.1214/23-STS898
© Institute of Mathematical Statistics, 2024

ANOVA for Metric Spaces, with Applications to Spatial Data

Raoul Müller, Dominic Schuhmacher and Jorge Mateu

Abstract. We give a review of some recent ANOVA-like procedures for testing group differences based on data in a metric space and present a new such procedure. Our statistic is derived from the classic Levene's test for detecting differences in dispersion. It uses only pairwise distances of data points and can be computed quickly and precisely in situations where the computation of barycenters ("generalized means") in the data space is slow, only by approximation or even infeasible. It also satisfies asymptotic normality.

We discuss the relative merits of the various procedures based on simulation studies for spatial point patterns and image data in a 1-way ANOVA setting. As applications, we perform 1- and 2-way ANOVAs on a data set of bubbles in a mineral flotation process and a data set of local pest counts in Madrid.

Key words and phrases: ANOVA, images, Levene's test, metric spaces, spatial point patterns.

1. INTRODUCTION

Real-world statistical data is often not Euclidean, involving components that are most suitably analyzed in a more complicated space. Examples include spaces of point patterns and more general subsets, trees and more general graphs, functions and images.

In recent years, a number of methods have been proposed for analyzing group differences of such data by generalizing classical analysis of variance (ANOVA) ideas to more complex data spaces. Examples include [12] for functional data, [30] for data on Riemannian manifolds and [36] for point pattern data. A common feature of the underlying spaces is that there is typically a more or less natural concept of distance between data points available. In addition to the more obvious choices of distances on function spaces and Riemannian manifolds, suitable metrics for tree spaces, graph spaces and point pattern spaces can be found in [7, 20] and [35], respectively.

In the present paper, we focus on generalized ANOVA procedures for metric spaces without using any more special structure of the space. There is a number of preceding articles that work in similar generality.

[2] proposes to perform ANOVA based on pairwise dissimilarities of observations rather than Euclidean distances between observations and their group means, and introduces the name PERMANOVA for this procedure. While not directly referring to any more abstract spaces than $\mathbb{R}^d$, that article clearly discusses the abstract template of doing non-Euclidean ANOVA without using a centroid object. We discuss this further in Section 3.1. [3] proposes multidimensional scaling followed by a Levene's test (using the centroid object in the principal coordinate space) for detecting differences of within-group dispersions (scatter, variability). This is referred to as PERMDISP; see [4]. [5] and [24] correct the PERMANOVA statistic for heteroscedasticity in the unbalanced setting based on the variants of classical ANOVA by Brown–Forsythe and Welch, respectively.

A somewhat different approach to decomposing an overall dispersion was introduced in [37] under the name of DISCO, but concentrated on αth powers of Euclidean distances, where $\alpha \in (0, 2]$, including classic ANOVA for the choice $\alpha = 2$.

More fundamentally different is the approach by [16], where an ANOVA procedure on metric spaces is designed using Fréchet means as centroid objects. The authors propose to use as statistic the sum of an ANOVA term and a Levene term. We discuss this further in Section 3.2.

Other methods, which we do not consider here, are graph-based approaches such as [45, 40], which use statistics based on between- and within-group counts of edges

Raoul Müller is a Research Assistant, Institute for Mathematical Stochastics, University of Göttingen, 37077 Göttingen, Germany (e-mail: raoul.mueller@uni-goettingen.de). Dominic Schuhmacher is a Professor, Institute for Mathematical Stochastics, University of Göttingen, 37077 Göttingen, Germany. Jorge Mateu is a Professor, Department of Mathematics, University Jaume I, 12071 Castellón, Spain.

in minimal spanning trees and possibly other distance-based graphs. These methods are typically designed for big data settings and do not directly take into account the set of all pairwise distances between observations.

Yet another approach comes from kernel-based methods. Very recently, [44] introduced a multisample generalization of the two-sample kernel method in [23], including various substantial improvements. In these methods, data elements on a separable metric space are mapped in a suitable reproducing kernel Hilbert space (RKHS) via the canonical feature map. A test for the equality of distributions is then based on the mean embeddings of the groups in the RKHS using an approximation of the asymptotic distribution under the null hypothesis. Kernel-based methods are especially suited for high-dimensional and functional data.

For further literature that largely falls into the above categories, we refer to the introductory sections of [44] and [40].

In the present paper, we formulate Anderson's PERMANOVA on general metric spaces. We simply refer to the resulting method as Anderson ANOVA, because the use of M (due to the use of $\mathbb{R}^d$ in Anderson's work) seems inappropriate in our context and the use of PER (referring to the fact that a permutation test is performed) does not distinguish it from the other methods used. Rather than pursuing the PERMDISP method mentioned above, we introduce a new test for detecting differences of within-group dispersion based on Levene's procedure and refer to it as L-test. Our test statistic works directly with the pairwise distances between observations without using any kind of group centroid, neither in the original metric space nor in any principal coordinate space. We show that it has an asymptotic χ_1^2-distribution, but we recommend using it with a permutation test just as the other statistics.

We also study the two summands used by [16] as separate test statistics for detecting differences in location and dispersion, respectively.

We refer to Table 1 for an overview of the methods discussed. This table does not necessarily provide a useful classification for the other approaches mentioned above. It is meant as a helpful guiding principle for the methods treated in more detail in the current paper and we use it as a starting point for the final discussion in Section 8.

TABLE 1
Overview of the non-Euclidean ANOVA methods studied in this paper. Procedures targeting location *are derived from the classic ANOVA statistic, whereas those targeting* dispersion *are derived from the classic Levene's statistic* (*ANOVA statistic for "deviations"*). *The rows distinguish whether computationally a procedure is based on* (*simple arithmetics of*) pairwise distances *or on a centroid object* (*here a* Fréchet mean) *in the metric space*

	Location	*Dispersion*
Pairwise distances	Anderson, Section 3.1	New L-test, Section 4
Fréchet means	Dubey–Müller, Section 3.2	Dubey–Müller, Section 3.2

Although the methods described are applicable in general metric spaces, our central goal in undertaking this research was to be able to perform ANOVA for point patterns and other spatial data; see also the discussion section of [35]. We therefore focus in the later part of the present paper on spaces of finite point patterns equipped with the TT-metric from [35] and space of images with an unbalanced or balanced Wasserstein metrics. As in many other spaces, exact Fréchet means can be computed within reasonable time only for (very) small data sets and one typically has to resort to a heuristic algorithm that finds only local minima of the Fréchet functional, which in our situations have been empirically investigated and are of good quality. We present simulation studies to compare the powers of the four tests across various situations and to understand the quality of approximation by the limiting χ_1^2-distribution from a practical point of view. We also present application of 1- and 2-way ANOVAs to two data sets: bubbles in a mineral flotation process and pest counts in the city of Madrid.

The plan of the paper is as follows: Section 2 contains a brief reminder of central aspects of classical ANOVA including Levene's test. In Section 3, we give a rather detailed presentation of Anderson ANOVA in metric spaces and the two summands proposed by Dubey and Müller. In Section 4, we introduce our new L-statistic, discuss its relation to the other methods and the original Levene's test, and show its asymptotic distribution. Section 5 is a short overview of the metric space of point patterns and the (unbalanced) Wasserstein space of images. In Sections 6 and 7, we present the simulation studies and the real-world data example. The paper ends with some further conclusions in Section 8.

2. CLASSIC ANOVA

For self-containedness and easy reference, we briefly remind the reader of some facts and formulae in the context of the classical ANOVA going back to [17]. Details can be found in [38].

One-way ANOVA. Given independent observations $x_{ij} \in \mathbb{R}$, $1 \leq j \leq n_i$, $1 \leq i \leq k$, from k potentially different distributions $P_1, \dots, P_k$, we do the following sum-of-squares decomposition

$$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS},$$

where

$$\mathrm{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 \quad \textit{(total sum of squares)},$$

$$\mathrm{MSS} = \sum_{i=1}^{k} n_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \quad \textit{(model sum of squares)},$$

$$\mathrm{RSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2 \quad \textit{(residual sum of squares)}.$$

Here, $\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ denotes the ith group mean and $\bar{x}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$ denotes the overall mean. We write $n = \sum_{i=1}^{k} n_i$ for the total number of observations.

Assume for now that the group distributions P_i are Gaussian with the same variance. Under the null hypothesis that also the means are the same (hence all data comes from the same normal distribution), it is well known that the ANOVA statistics

$$F = \frac{n-k}{k-1} \frac{\mathrm{MSS}}{\mathrm{RSS}}, \tag{1}$$

describing the ratio between the variability explained by the model and the total variability in the data, is F-distributed with $k-1$ and $n-k$ degrees of freedom. Since $T \sim F(d_1, d_2)$ implies $d_1 T \xrightarrow{\mathcal{D}} \chi^2_{d_1}$ as $d_2 \to \infty$, we obtain

$$(k-1)F \xrightarrow{\mathcal{D}} \chi^2_{k-1} \quad \text{as } n \to \infty. \tag{2}$$

The *asymptotic* result remains true even if $P_1 = P_2 = \cdots = P_k$ is non-Gaussian, but has second moments and there are $\lambda_1, \dots, \lambda_k > 0$ such that the ratios of group sizes satisfy $\frac{n_i}{n} \to \lambda_i$; see, for example, [43], Section 3.6.2.

REMARK 1. Strictly speaking ANOVA techniques are designed for inference within a linear model of different group means plus errors. Based on an error distribution P with mean zero, one considers the model equations

$$x_{ij} = \mu_i + \varepsilon_{ij}, \quad 1 \le j \le n_i, 1 \le i \le k,$$

where $\mu_i \in \mathbb{R}$ are the different group means and ε_{ij} are i.i.d. P-distributed error terms. In terms of the group distributions above this means that $P_i = P * \delta_{\mu_i}$, that is, P_i is obtained by shifting P by μ_i. Note that the asymptotic χ^2_{k-1}-test does not need this assumption since in any case the null hypothesis just correspond to having k times the same distribution. At the same time, we cannot expect this test to achieve high power against all alternatives that have substantially different group distributions (see also the paragraph on Levene's test below). We will take up this point when discussing ANOVA on metric spaces, where typically "shifting the distribution" is meaningless (but may have an intuitive counterpart).

Two-way ANOVA. As soon as more than one grouping factor is involved, important design decisions come into play, such as if factors are (partially) nested or if we allow for interaction terms between several factors on the same level. ANOVA has a long standing history with many different designs. As an example, which is pursued further in later sections, we remind the reader of the balanced two-way ANOVA (two main factors, with interaction terms, same number $\tilde{n}$ of observations for each factor combination).

Given independent observations $x_{i_1 i_2 j} \in \mathbb{R}$, $1 \le j \le \tilde{n}$, $1 \le i_1 \le k_1$, $1 \le i_2 \le k_2$ from groups obtained by crossing a Factor a with k_1 levels and a Factor b with k_2 levels (with $n_{i_1 i_2} := \tilde{n}$ observations for each combination), we can perform a finer sum-of-squares decomposition

$$\mathrm{TSS} = \mathrm{SSa} + \mathrm{SSb} + \mathrm{SSi} + \mathrm{RSS},$$

splitting up the model sum of squares into sums of squares for the individual factors and an interaction sum of squares. In formulae,

$$\mathrm{TSS} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{\tilde{n}} (x_{i_1 i_2 j} - \bar{x}_{\dots})^2,$$

$$\mathrm{RSS} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{\tilde{n}} (x_{i_1 i_2 j} - \bar{x}_{i_1 i_2 \cdot})^2,$$

$$\mathrm{SSa} = \sum_{i_1=1}^{k_1} k_2 \tilde{n} (\bar{x}_{i_1 \cdot\cdot} - \bar{x}_{\dots})^2,$$

$$\mathrm{SSb} = \sum_{i_2=1}^{k_2} k_1 \tilde{n} (\bar{x}_{\cdot i_2 \cdot} - \bar{x}_{\dots})^2,$$

$$\mathrm{SSi} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \tilde{n} (\bar{x}_{i_1 i_2 \cdot} - \bar{x}_{i_1 \cdot\cdot} - \bar{x}_{\cdot i_2 \cdot} + \bar{x}_{\dots})^2,$$

where the various means are taken over the dot components while keeping the given indices fixed. Set $n = k_1 k_2 \tilde{n} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} n_{i_1 i_2}$.

In addition to performing an omnibus test for group differences as for one-way ANOVA, we may then test for effects of Factors a and b separately, as well as for an interaction effect. The corresponding statistics are

$$Fa = \frac{n - k_1 k_2}{k_1 - 1} \frac{\mathrm{SSa}}{\mathrm{RSS}}, \qquad Fb = \frac{n - k_1 k_2}{k_2 - 1} \frac{\mathrm{SSb}}{\mathrm{RSS}},$$

$$Fi = \frac{n - k_1 k_2}{(k_1 - 1)(k_2 - 1)} \frac{\mathrm{SSi}}{\mathrm{RSS}}.$$

If the observations come from Gaussian distributions with equal variances, each of the three statistics is F-distributed again under the corresponding null hypothesis that different levels of the factor or interaction to be tested do not lead to different shifts in mean. The degrees of freedom can be read from the denominator and the numerator, respectively, of the first ratio in each statistic.

Levene's test. The test first proposed in [32] was originally developed as a preliminary test to check for equal variances *before* applying the basic ANOVA F-test in the Gaussian setting. This was important, as it was well known at the time that for the goal of inference about differences in the means of the various groups (see Remark 1), the size of the F-test can depart substantially from its nominal size if group variances are not equal.

[32] proposed to use as test statistic the usual ANOVA statistic, but to replace the observations x_{ij} by the absolute differences from their group means $z_{ij} = |x_{ij} - \bar{x}_{i\cdot}|$, that is,

$$\widetilde{F} = \frac{n-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i\cdot})^2}. \tag{3}$$

If the observations are independently sampled from the same Gaussian distributions, it is plausible that $\widetilde{F}$ is still approximately F-distributed, because the dependence between the z_{ij} is small even at moderate group sizes. This was confirmed by simulation in [32]. [10] present a larger simulation experiment suggesting that replacing the $\bar{x}_{i\cdot}$ in the definition of z_{ij} by a trimmed mean or median leads to a more robust test for non-Gaussian data.

Current best practice suggests to perform a Welch-modified ANOVA directly if the assumption of equal variance is unclear as it results only in a small loss of power in the case where the variances are indeed equal. We refer to [18] for a comprehensive presentation on Levene's test including this question and many further developments.

Levene's test and its variants remain highly important today as differences in variances (or some other measure of dispersion) are often in the center of attention in their own rights. In the rest of the paper, we present tests on differences in "location" of groups and differences in "dispersion" of groups, both based on interpoint distances in a metric space. Our goal is to combine one of either kind in order to detect group differences in some universality.

3. NON-EUCLIDEAN ANOVA

In this and the next sections, we assume that our data lies in a general metric space $(\mathcal{X}, d)$. We present existing methods of testing for group differences based on ANOVA-like ideas. For the presentation, we focus on generalizations of 1-way ANOVA, but provide further information on which methods can easily be extended to more complex designs. We always assume having $n = \sum_{i=1}^{k} n_i$ independent observations $x_{ij} \in \mathcal{X}$, $1 \le j \le n_i$, $1 \le i \le k$ from k potentially different distributions $P_1, \dots, P_k$ on $\mathcal{X}$ (with Borel σ-algebra).

3.1 Anderson ANOVA

[2] argues, in the context of data sets in ecology, that traditional multivariate analogues of ANOVA are too stringent in their assumptions. These are typically based on similar statistics as (1), but with absolute values replaced by Euclidean norms; see, for example, [33] Section 12.3. We may avoid the use of means of observations by writing TSS − RSS instead of MSS and replacing the sums of squared deviations from the mean with the help of the formula

$$\begin{aligned}\sum_{j=1}^{m} \|y_j - \overline{y}\|^2 &= \frac{1}{2m} \sum_{j_1, j_2=1}^{m} \|y_{j_1} - y_{j_2}\|^2 \\ &= \frac{1}{m} \sum_{j_1, j_2=1}^{m,<} \|y_{j_1} - y_{j_2}\|^2,\end{aligned}$$

where we indicate by "<" in the summation bound that the sum is to be taken over strictly ordered summands only; here, $j_1 < j_2$. Anderson proposes to replace the pairwise Euclidean distances by more general dissimilarities between observations and performs a permutation test. In our context, we simply use the pairwise distances in the metric space. Thus,

$$\begin{aligned}\mathrm{TSS} &= \frac{1}{n}\Bigg(\sum_{i_1, i_2=1}^{k,<} \sum_{j_1=1}^{n_{i_1}} \sum_{j_2=1}^{n_{i_2}} d^2(x_{i_1 j_1}, x_{i_2 j_2}) \\ &\qquad + \sum_{i=1}^{k} \sum_{j_1, j_2=1}^{n_i,<} d^2(x_{i j_1}, x_{i j_2}) \Bigg), \\ \mathrm{RSS} &= \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j_1, j_2=1}^{n_i,<} d^2(x_{i j_1}, x_{i j_2}), \\ \mathrm{MSS} &= \mathrm{TSS} - \mathrm{RSS}\end{aligned}$$

and the final Anderson ANOVA statistic becomes

$$F_{\mathrm{A}} = \frac{n-k}{k-1} \frac{\mathrm{MSS}}{\mathrm{RSS}}.$$

It has been noted in various places that this statistic may suffer from type I error inflation (in terms of a null hypothesis of equal *means* in Euclidean space) and substantial loss of power in the unbalanced setting if the groups are heteroscedastic; see, for example, [1]. [5] and [24] propose improvements based on the classical ANOVA variants by [11] and [42], respectively. In the former, the F-statistic is replaced by

$$F_{\mathrm{BF}} = \frac{\mathrm{MSS}}{\sum_{i=1}^{k} (1 - \frac{n_i}{n}) \frac{1}{n_i(n_i-1)} \sum_{j_1, j_2=1}^{n_i,<} d^2(x_{i j_1}, x_{i j_2})}.$$

For the simulation studies in Section 6, we concentrate on the balanced setting, for which Anderson F_A performs typically well even in presence of heteroscedacity. We therefore do not discuss these improvements further, which in the balanced setting do not change the statistic.

3.2 Fréchet ANOVA

[16] introduce ANOVA-like terms that use distances in the metric d to Fréchet means rather than absolute differences to averages. For observation $y_1, \dots y_m \in \mathcal{X}$, the

Fréchet mean is defined as

$$\bar{y} = \operatorname*{argmin}_{z \in \mathcal{X}} \sum_{i=1}^{m} d^2(y_i, z). \tag{4}$$

One of the assumptions in [16] is that all Fréchet means considered exist and are unique. For our usual set of observations, we denote by $\bar{x}_{i\cdot}$ the Fréchet mean of $x_{i1}, \dots, x_{in_i}$, $i = 1, \dots, k$ and by $\bar{x}_{\cdot\cdot}$ the Fréchet mean of all observations. Following the notation in [16], we write the Fréchet variance for the ith group and the total Fréchet variance as

$$\hat{V}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_{ij}, \bar{x}_{i\cdot}) \quad \text{and}$$

$$\hat{V}_p = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} d^2(x_{ij}, \bar{x}_{\cdot\cdot}),$$

respectively. While $\hat{V}_i$ is the mean of $d^2(x_{ij}, \bar{x}_{i\cdot})$, $j = 1, \dots, n_i$, we also require the corresponding variance

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} d^4(x_{ij}, \bar{x}_{i\cdot}) - \hat{V}_i^2.$$

Setting $\lambda_i = \frac{n_i}{n}$, one finally obtains

$$U_n = \sum_{i_1, i_2 = 1}^{k, <} \frac{\lambda_{i_1} \lambda_{i_2}}{\hat{\sigma}_{i_1}^2 \hat{\sigma}_{i_2}^2} (\hat{V}_{i_1} - \hat{V}_{i_2})^2,$$

$$F_n = \hat{V}_p - \sum_{i=1}^{k} \lambda_i \hat{V}_i,$$

$$T = \frac{n U_n}{\sum_{i=1}^{k} \frac{\lambda_i}{\hat{\sigma}_i^2}} + \frac{n F_n^2}{\sum_{i=1}^{k} \lambda_i^2 \hat{\sigma}_i^2} =: T_L + T_F.$$

In the Euclidean setting of Section 2, the term F_n is equal to $\frac{1}{n}(\text{TSS} - \text{RSS})$ and the denominator of T_F is then an estimator for the variance of $\frac{1}{n}\text{RSS}$, so that T_F has close ties to the ANOVA F-statistic. The unweighted summands $(\hat{V}_{i_1} - \hat{V}_{i_2})^2$ of U_n are similar in spirit to the terms $(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2$ from the definition of Levene's statistic, and in fact it appears that in the Euclidean case T_L corresponds exactly to a simpler variant of Welch's ANOVA applied to $d^2(x_{ij}, \bar{x}_{i\cdot})$, $j = 1, \dots, n_i$, $i = 1, \dots, k$; see the computation in formulae (8)–(16) in [24]. Thus, T_L has close ties to Levene's statistic.

Dubey and Müller show under a list of conditions pertaining to existence and uniqueness of theoretical and empirical Fréchet means and the complexity of the metric space (in terms of entropy integrals) that

$$\frac{n U_n}{\sum_{i=1}^{k} \frac{\lambda_i}{\hat{\sigma}_i^2}} \xrightarrow{\mathcal{D}} \chi_{k-1}^2 \quad \text{and}$$

$$\frac{n F_n^2}{\sum_{i=1}^{k} \lambda_i^2 \hat{\sigma}_i^2} \xrightarrow{\mathcal{D}} 0 \quad \text{as } n \to \infty.$$

The authors advocate the simple addition of the two terms in order to obtain a single test statistic T, maybe with weights if there is prior information available whether to rather look out for inequality of Fréchet means or of Fréchet variances. However, due to the unbalanced convergence of the two terms and the fact that the reason for the concrete normalization (especially) of T_F remains a bit inscrutable to us, we prefer to analyze the two summands separately in Section 6.

4. A NEW NON-EUCLIDEAN METHOD OF LEVENE TYPE

What appears to be missing is a test for detecting differences of within-group dispersion that is based directly on pairwise distances between observations in the metric space. The idea of the PERMDISP test mentioned in the Introduction, that is, performing multidimensional scaling and applying Levene's test in the principal coordinate space, is to some extent applicable here. However, it is rather an indirect method and it is methodologically not on the same level as the Anderson F_A. Indeed multidimensional scaling can be applied in combination with *any* Euclidean procedure, so the PERMDISP method should be rather paired up with the analog method of multidimensional scaling plus applying Euclidean (M)ANOVA. What is more, it contains an unwelcome tuning parameter, the number of principal coordinates, which is not easy to choose, but may be crucial. Instead we propose the following test of Levene type for data in a metric space.

4.1 Form and Properties

We assume the same setup as in the previous section, that is, there are $n = \sum_{i=1}^{k} n_i$ independent observations $x_{ij} \in \mathcal{X}$, $1 \le j \le n_i$, $1 \le i \le k$ from k potentially different distributions $P_1, \dots, P_k$ on $\mathcal{X}$. Set $N_i = \binom{n_i}{2}$ and $N = \sum_{i=1}^{k} N_i$. As a surrogate for the individual deviation terms z_{ij} from Levene's statistic (3), which in a general metric space would require the use of a Fréchet or similar mean, we use $d_{i,\{j_1,j_2\}} := \frac{1}{2} d(x_{ij_1}, x_{ij_2})$. To simplify the notation, we enumerate the two-element subsets of $\{1, \dots, n_i\}$ by $j = 1, \dots, N_i$ and use d_{ij} rather than $d_{i,\{j_1,j_2\}}$ for the jth half-distance in the ith group.

In a first step, we assume that $n_1 = \cdots = n_k$ (balanced case) and emulate the statistics (3) by setting

$$L := \frac{N - k}{k - 1} \frac{\sum_{i=1}^{k} n_i (\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (d_{ij} - \bar{d}_{i\cdot})^2}, \tag{5}$$

where

$$\bar{d}_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} d_{ij} \quad \text{and} \quad \bar{d}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} d_{ij}$$

denote the ith group mean and the overall mean over pairwise distances, respectively.

Typographically, the main fractions of equations (5) and (3) are very similar, but the way they use the data x_{ij} is quite different in that we replace $z_{ij} = |x_{ij} - \bar{x}_{i\cdot}|$, $1 \le j \le n_i$ by $d_{i,\{j_1,j_2\}} = \frac{1}{2}d(x_{ij_1}, x_{ij_2})$, $1 \le j_1 < j_2 \le n_i$. Note that we keep n_i in the numerator rather than replacing it by N_i, which might have seemed more natural at first glance. The reason is the substantial dependence of the random variables $d_{i,\{j_1,j_2\}}$ (as opposed to the less substantial dependence between the z_{ij}) for each i, which implies that n_i, not N_i, is the correct scaling factor; see Section 4.2. Note further that, for the same reason, the main denominator is not the most natural choice here, but it is convenient since it keeps the statistic similar to the original Levene statistic, is fast to compute and empirically performs no worse than the more natural choice discussed in Section 4.2.

Note that for a translation invariant metric on a group (in the algebraic sense of the word) we may perform joint translations of all data points within each of the k observation groups separately without changing the value of L. As a consequence, the statistic L cannot detect differences between $P_1, \dots, P_k$ that are purely due to different locations. The same holds true for the classic Levene's statistic and the statistics introduced in (6) and (7) below.

There are various ways how one might generalize (5) to general group sizes. We propose using

$$(6)\qquad L := \frac{N-k}{k-1}\frac{\frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (\bar{d}_{i\cdot} - \bar{d}_{j\cdot})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(d_{ij} - \bar{d}_{i\cdot})^2}.$$

Direct computation shows that equations (6) and (5) agree in the balanced case, but not in general; see Remark 9. The statistic (6) performs well in several respects: it allows for an asymptotic distribution (χ^2_{k-1} up to a deterministic factor, see Corollary 3), is still fast to compute and shows a reasonable performance with unequal group sizes (see also Section 6.1), though it may well be that a more judicious scaling that takes more proper care of different group sizes would be superior.

We briefly come back to this last point in Section 6, but do not go much deeper in the present paper because based on additional considerations, both theoretical and from simulation studies, we do not see any clear improvements when choosing different normalizations.

In spite of the limit distribution, which we compute in the next section, we recommend performing a permutation test as for the other statistics considered. For this, we permute the observations, not only their distances, that is, new permutations use distances that are potentially different from the pairwise within-group distances of the original data. As a consequence, not only the RSS changes with permutations, but also the TSS.

It is easy enough to generalize the construction of the above test statistic to more complex experimental designs. As an example, we take up the balanced two-way ANOVA from Section 2 and form the corresponding Levene-type statistics for $(\mathcal{X}, d)$. For the specific statistics, see Section 7.1.

4.2 Limit Distribution

In this subsection, we derive asymptotic distributions for the statistic L from (6) and for the related statistic

$$(7)\quad \widetilde{L} := \frac{N^* - k}{k-1}\frac{\frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (\bar{d}_{i\cdot} - \bar{d}_{j\cdot})^2}{4T_n},$$

where $N^* = \sum_{i=1}^{k} n_i(n_i - 1)^2$ and

$$(8)\quad T_n = \sum_{i=1}^{k}\sum_{\substack{j_1,j_2,j_3=1\\ j_1 \notin \{j_2,j_3\}}}^{n_i} (d_{i,\{j_1,j_2\}} - \bar{d}_{i\cdot})(d_{i,\{j_1,j_3\}} - \bar{d}_{i\cdot}).$$

The previous formula makes it necessary to use the more complicated notation $d_{i,\{j_1,j_2\}} = \frac{1}{2}d(x_{ij_1}, x_{ij_2})$ from the beginning of Section 4.1. Note that $\frac{1}{N^*-k}T_n$ is a natural group based estimator of $\mathrm{Cov}(\frac{1}{2}d(X_1, X_2), \frac{1}{2}d(X_1, X_3))$, where X_1, X_2, X_3 are three independent random variables sampled from the distribution of the group. The normalization by $N^* - k$ rather than N^* is simply modeled after the bias correcting term for independent data points.

In spite of the ANOVA-like construction, we cannot use the asymptotic theory for ANOVA directly, because the distances $d_{i,\{j_1,j_2\}}$, our "data," stem from dependent random variables for each i. This dependence is taken into account by using the factor $\frac{n_i n_j}{n}$ rather than N_i or N_j in the numerator and by normalizing with $\frac{1}{N^*-k}4T_n$ in (7), which then still allows to obtain the asymptotic χ^2_{k-1}-distribution for $(k-1)\widetilde{L}$. In contrast, $(k-1)L$ converges "only" toward a multiple of χ^2_{k-1} that depends on parameters of the group distribution.

THEOREM 2. *Assume that the Borel σ-algebra for $(\mathcal{X}, d)$ is countably generated. In the usual* 1*-way setup of Section* 4.1, *assume that $P_1 = \cdots = P_k = P$ for a distribution P that is not a Dirac distribution and satisfies $\int_{\mathcal{X}}\int_{\mathcal{X}} d^2(x, y)P(dx)P(dy) < \infty$. Suppose that there are $\lambda_i > 0$ such that $n_i/n \to \lambda_i$ for every i as $n \to \infty$. Then we have*

$$(k-1)\widetilde{L} \xrightarrow{\mathcal{D}} \chi^2_{k-1} \quad \textit{as } n \to \infty.$$

COROLLARY 3. *Under the conditions of Theorem* 2, *we obtain*

$$(k-1)L \xrightarrow{\mathcal{D}} \frac{4\gamma^2}{\sigma^2}\chi^2_{k-1} \quad \textit{as } n \to \infty,$$

where with independent $X, Y, Z \sim P$ we have

$$\gamma^2 = \mathrm{Cov}\big(d(X, Y), d(X, Z)\big);$$

$$\sigma^2 = \mathrm{Var}\big(d(X, Y)\big).$$

Proof of Theorem 2. Under the null hypothesis, our data is generated by independent $\mathcal{X}$-valued random elements $X_{ij} \sim P$, $1 \le j \le n_i$, $1 \le i \le k$ and the distances $d_{i,\{j_1,j_2\}}$ are realizations of the random variables $\frac{1}{2}d(X_{ij_1}, X_{ij_2})$, $1 \le j_1 < j_2 \le n_i$, $1 \le i \le k$. Under the conditions on P, we have asymptotic normality of the U-statistics

$$U_i = U_i^{(n)} = \binom{n_i}{2}^{-1} \sum_{j_1,j_2=1}^{n_i,<} \frac{1}{2} d(X_{ij_1}, X_{ij_2}), \quad \text{for } i = 1, \dots, k \tag{9}$$

by a straightforward generalization of Hoeffding's theorem to random elements in $\mathcal{X}$; see Theorem 6 in the Appendix. More precisely, we have with $X, Y, Z \sim P$ independent that

$$\sqrt{n_i}\Big(U_i - \frac{1}{2}\mathbb{E}d(X,Y)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \gamma^2) \quad \text{as } n_i \to \infty, \tag{10}$$

where $\gamma^2 = \operatorname{Cov}(d(X,Y), d(X,Z)) = \operatorname{Var}(\mathbb{E}(d(X,Y) | X)) = 4\gamma_h^2$ in the notation of the Appendix with $h = \frac{1}{2}d$. In view of the 1-way ANOVA construction, on which L is based, we define the "design matrix" $D = D_n \in \mathbb{R}^{n \times k}$ by

$$D' := \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \dots & 1 & \cdots & 0 & \cdots & 0 \\ & \vdots & & & & & \ddots & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{k \times n}, \tag{11}$$

where the ith row has exactly n_i ones, and the "contrast matrix"

$$C := \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & & 0 & -1 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(k-1) \times k}. \tag{12}$$

Setting $\Delta = \lim_{n\to\infty} \frac{1}{n} D_n' D_n = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, we obtain with $U = U^{(n)} = (U_1, \dots, U_k)'$ by independence of the components and $n_i \to \infty$ as $n \to \infty$ (since $\lambda_i > 0$) that

$$Z_n := \gamma^{-1}\sqrt{n}\Delta^{1/2}(U - \mathbb{E}U) \xrightarrow{\mathcal{D}} \mathcal{N}_k(0, I_k) \quad \text{as } n \to \infty. \tag{13}$$

Setting $\nu = (n_1, \dots, n_k)'$, we may further compute $C'(C(D'D)^{-1}C')^{-1}C = D'D - \frac{1}{n}\nu\nu'$ (see Lemma 8 in the Appendix for the calculation) and, therefore,

$$\widetilde{L} = \frac{N^* - k}{k-1} \frac{U'C'(C(D_n'D_n)^{-1}C')^{-1}CU}{4T_n}. \tag{14}$$

Since $\mathbb{E}U = \frac{1}{2}\mathbb{E}d(X,Y) \cdot \mathbb{1} \in \mathbb{R}^k$ and $C \cdot \mathbb{1} = 0$, we obtain

$$(k-1)\widetilde{L} = \gamma^2 \frac{Z_n'(\frac{1}{n}W_n)Z_n}{\frac{4}{N^*-k}T_n},$$

where $W_n := \Delta^{-1/2}C'(C(D_n'D_n)^{-1}C')^{-1}C\Delta^{-1/2}$. Note that

$$W := \lim_{n\to\infty} \frac{1}{n} W_n = \Delta^{-1/2}C'\big(C\Delta^{-1}C'\big)^{-1}C\Delta^{-1/2}$$

is a symmetric and idempotent matrix of rank $k-1$ and, therefore, $Z'WZ \sim \chi^2_{k-1}$ for $Z \sim \mathcal{N}_k(0, I_k)$ by Lemma 10 from the Appendix. Using (13), it is straightforward to show with the help of the continuous mapping theorem that

$$Z_n'\Big(\frac{1}{n}W_n\Big)Z_n \xrightarrow{\mathcal{D}} \chi^2_{k-1}.$$

So, it suffices to show that $\frac{1}{N^*-k}T_n \xrightarrow{p} \gamma^2_{d/2}$. For this, we note that the normalized inner sum of (8) satisfies

$$\begin{aligned} &\frac{1}{n_i(n_i-1)^2} \sum_{\substack{j_1,j_2,j_3=1 \\ j_1 \notin \{j_2,j_3\}}}^{n_i} (d_{i,\{j_1,j_2\}} - \bar{d}_{i\cdot})(d_{i,\{j_1,j_3\}} - \bar{d}_{i\cdot}) \\ &\quad = \underbrace{\frac{n_i(n_i-1)(n_i-2)}{n_i(n_i-1)^2}}_{\longrightarrow 1} \\ &\qquad \times \underbrace{\frac{1}{n_i(n_i-1)(n_i-2)} \sum_{j_1,j_2,j_3=1}^{n_i,\neq} (d_{i,\{j_1,j_2\}} - \bar{d}_{i\cdot})(d_{i,\{j_1,j_3\}} - \bar{d}_{i\cdot})}_{\longrightarrow \operatorname{Cov}(\frac{1}{2}d(X_1,X_2), \frac{1}{2}d(X_1,X_3)) = \gamma^2_{d/2}} \\ &\qquad + \underbrace{\frac{1}{n_i-1}}_{\longrightarrow 0} \underbrace{\frac{1}{n_i(n_i-1)} \sum_{j_1,j_2}^{n_i,\neq} (d_{i,\{j_1,j_2\}} - \bar{d}_{i\cdot})^2}_{\longrightarrow \operatorname{Var}(\frac{1}{2}d(X_1,X_2)) = \sigma^2_{d/2}}, \end{aligned} \tag{15}$$

where convergence of the averages is almost surely and follows after expansion of the products by the strong law of large numbers for U-statistics using the prerequisite $\mathbb{E}(d(X_1,X_2)^2) < \infty$; see [29].

Thus, for the total term,

$$\begin{aligned} &\frac{1}{N^*-k} T_n \\ &\quad = \frac{1}{N^*-k} \sum_{i=1}^{k} n_i(n_i-1)^2 \cdot \frac{1}{n_i(n_i-1)^2} \\ &\qquad \times \sum_{\substack{j_1,j_2,j_3=1 \\ j_1 \notin \{j_2,j_3\}}}^{n_i} (d_{i,\{j_1,j_2\}} - \bar{d}_{i\cdot})(d_{i,\{j_1,j_3\}} - \bar{d}_{i\cdot}) \\ &\quad \longrightarrow \gamma^2_{d/2}. \end{aligned}$$

□

Proof of Corollary 3. This follows from Theorem 2 because

$$L = 4 \frac{\frac{1}{N^*-k}T_n}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(d_{ij} - \bar{d}_{i\cdot})^2} \widetilde{L},$$

where the numerator is a consistent estimator of $\gamma^2/4$ and the denominator is a consistent estimator of $\sigma^2/4$; see (15). □

REMARK 4. The convergence (13) remains true under the alternative hypothesis if we replace γ by a $k \times k$ diagonal matrix with entries $\gamma_i = \mathrm{Cov}_{P_i}(d(X,Y), d(X,Z))$, $1 \le i \le k$. This opens up the way for studying the distribution of $\widetilde{L}$ under certain alternatives. For example, assuming for simplicity again that $\gamma_i = \gamma$ for all i, we obtain under an alternative of the form

$$H_A : \mathbb{E}_{P_i} d(X,Y) = \mu + n^{-1/2} h_i, \quad 1 \le i \le k,$$

for some $\mu \in \mathbb{R}$ and some $h \in \mathbb{R}^k \setminus \{0\}$ the convergence

$$Z_n := \gamma^{-1}\sqrt{n}\Delta^{1/2}(U - \mu\mathbb{1}) \xrightarrow{\mathcal{D}} \mathcal{N}_k(\gamma^{-1}\Delta^{1/2}h, I_k)$$

as a replacement of (13). Therefore, continuing in the analogous way as in the proof of Theorem 2 and using Lemma 10 in its general form, we obtain under H_A,

$$(k-1)\widetilde{L} \xrightarrow{\mathcal{D}} \chi^2_{k-1}(\gamma^{-2}h'\Delta h) \quad \text{as } n \to \infty,$$

that is, a noncentral χ^2-limit distribution with noncentrality parameter $\gamma^{-2}h'\Delta h$.

We assess the convergence speed of $\widetilde{L}$ under the null hypothesis in the context of simulated point pattern data in Appendix B. While convergence seems to be rather fast in the simulated examples, we use permutation-based tests in the remainder of the paper. This ensures an honest p-value across all the different examples, albeit one that is random because permutations must be sampled in all but the smallest of examples. Permutation tests are also in the tradition of previous methods; see, for example, [4, 16].

5. METRIC SPACES OF SPATIAL DATA

In Sections 6 and 7, we apply the four statistics from Table 1 to spatial data, most prominently point pattern data. For self-containedness, we give a short summary of the relevant concepts for the space of finite point patterns equipped with the metric introduced in [35], to which we refer as MSM20, as well as to the space of images equipped with an unbalanced Wasserstein metric.

5.1 Finite Point Pattern Data

For $n \in \mathbb{Z}_+$, write $[n] = \{1, 2, \dots, n\}$ (including $[0] = \varnothing$). Denote by $\mathfrak{N}_{\text{fin}}$ the space of finite multisets on a complete separable metric space $(\mathcal{R}, \varrho)$. We refer to the elements $\xi = \{x_1, x_2, \dots, x_n\} \in \mathfrak{N}_{\text{fin}}$ as point patterns, where $n \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ and $x_i \in \mathcal{X}$, $i \in [n]$. Note that $x_i = x_j$ for $i \ne j$ is allowed and that the point patterns can be identified with the counting measure $\sum_{i=1}^n \delta_{x_i}$, which is often helpful for theoretical considerations. We write $|\xi|$ to denote the total number of points in the pattern ξ.

DEFINITION 5 (Definition 1 of MSM20). Let $C > 0$ and $p \ge 1$ be two parameters, referred to as *penalty* and *order*, respectively.

For $\xi = \{x_1, \dots, x_m\}, \eta = \{y_1, \dots, y_n\} \in \mathfrak{N}_{\text{fin}}$ define the *transport-transform (TT) metric* by

$$\begin{aligned} d_{\mathrm{TT}}(\xi,\eta) &= d_{\mathrm{TT}}^{(C,p)}(\xi,\eta) \\ &= \biggl(\min\biggl((m+n-2l)C^p \\ &\qquad + \sum_{r=1}^{l} \varrho(x_{i_r}, y_{j_r})^p\biggr)\biggr)^{1/p}, \end{aligned} \tag{16}$$

where the minimum is taken over equal numbers of pairwise different indices $i_1, \dots, i_l$ in $[m]$ and $j_1, \dots, j_l$ in $[n]$, that is, over the set

$$\begin{aligned} S(m,n) = \bigl\{&(i_1, \dots, i_l; j_1, \dots, j_l); \\ &l \in \{0, 1, \dots, \min\{m,n\}\}, \\ &i_1, \dots, i_l \in [m] \text{ pairwise different}, \\ &j_1, \dots, j_l \in [n] \text{ pairwise different}\bigr\}. \end{aligned}$$

The distance $d_{\mathrm{TT}}(\xi, \eta)$ can be computed by filling up the smaller point pattern with dummy points located at distance C until it has the same cardinality n as the larger point pattern and then solving a standard assignment problem with cost $\min\{d(x,y), 2^{1/p}C\}$ between points x, y (MSM20, Theorem 1). The classical worst-case complexity of this is $O(n^3)$ (MSM20, Remark 1), which can be somewhat improved to order $n^{2.5}$ up to polylogarithmic factors [31]. Practical computation times for well over $n = 1000$ points are less than one second (R package ttbary, [34], using the auction algorithm from [6]).

The TT-metric can be interpreted as an unbalanced Wasserstein metric (MSM20, Remark 3). Computing Fréchet means in Wasserstein spaces is a topic of active research; see, for example, [8, 9, 27] and references therein for recent developments the space of discrete measures. In our context, an additional increase in difficulty comes from the constraint that the result must be a discrete measure with integer cardinality.

In MSM20, we therefore propose a heuristic to obtain a local minimizer of the Fréchet functional in (4), which we refer to as *pseudo-barycenter*. Our algorithm starts by filling up all the point patterns with dummy points to a suitable number n (which may have to be larger than the maximal cardinality of the point patterns) and initializing a pseudo-barycenter at random. Then it proceeds by alternating between matching the points of each pattern to the current pseudo-barycenter and recentering each point of the pseudo-barycenter within each cluster of matched points in a way similar to a k-means algorithm, but with some additional heuristics for switching pseudo-barycenter points between dummy points and real points. This procedure converges in finitely many steps.

The resulting pseudo-barycenters are obtained much faster and appear to be of good quality (consistent objective function values and results conform with intuition), but are by no means perfect and still require considerable computation time for hundreds of patterns with hundred of points (Tables 1–4 in MSM20). All of the Dubey–Müller statistics in Section 6.1 and 6.2 are based on such pseudo-barycenters.

In view of the conditions for Theorem 2, completeness and separability are inherited from $(\mathcal{X}, \varrho)$ to $(\mathfrak{N}_{\text{fin}}, d_{\text{TT}})$. This is straightforward to see by the fact that $d_{\text{TT}}(\xi_N, \xi) \to 0$ iff $|\xi_N| \to |\xi|$ and each point x of ξ is approximated by exactly one point of ξ_N (if x is a multipoint of cardinality k, this means that there is a total of k points in ξ_N, possibly forming multipoints of their own, that converge to x). Since $d_{\text{TT}}(\xi, \eta) \leq 2^{1/p} C \max\{|\xi|, |\eta|\}^{1/p}$, the condition $\int_{\mathcal{X}} \int_{\mathcal{X}} d^2(x,y) P(dx) P(dy) < \infty$ is satisfied for d_{TT} as long as $\mathbb{E}|\Xi|^{2/p} < \infty$ for $\Xi \sim P$, which is the case for all point process distributions considered in Section 6.

For the simulation study in the next section, it is helpful to understand some basic probability measures on $\mathfrak{N}_{\text{fin}}$. Suppose that $\mathcal{R} \subset \mathbb{R}^d$ is compact (in the next section we only use a unit square in $\mathbb{R}^2$). A random element in the metric space $(\mathfrak{N}_{\text{fin}}, d_{\text{TT}})$, equipped with its Borel σ-algebra is called a *point process*, that is, a point process is a measurable map from a probability space to $\mathfrak{N}_{\text{fin}}$. The Borel σ-algebra coincides with the smallest σ-algebra that makes $\xi \mapsto \xi(A)$ measurable for every measurable $A \subset \mathcal{R}$, which is the usual σ-algebra considered on $\mathfrak{N}_{\text{fin}}$; see Proposition 9.1.IV in [14].

We say a point process Ξ satisfies *complete spatial randomness (CSR)* if it is a Poisson process with intensity measure $\nu = \lambda \text{Leb}^d$, where $\lambda \geq 0$ and Leb^d is Lebesgue measure (on $\mathcal{R}$). This means that $\Xi(A) \sim \text{Po}(\nu(A))$ for all measurable $A \subset \mathcal{R}$ and that $\Xi(A_1), \dots, \Xi(A_l)$ are independent for all $l \in \mathbb{N}$ and all measurable $A_1, \dots, A_l \subset \mathcal{R}$ that are pairwise disjoint; see, for example, Section 2.4 in [13] for more details on the Poisson process.

5.2 Image Data

By an *image*, we mean here simply an $r \times s$ matrix with nonnegative entries, where $r, s \in \mathbb{N}$. Real data may come from satellite pictures, medical imaging methods or microscopy, among others. Given two images $X = (x_{i,j})$, $Y = (y_{k,l}) \in \mathbb{R}_+^{r \times s}$ and pixel distances $\rho_{(i,j),(k,l)}$ that come typically from an ℓ_1- or ℓ_2-metric, we may define an unbalanced Wasserstein metric of order $p \geq 1$ by

$$
\begin{aligned}
&d_{\text{UBW}}(X, Y) \\
&\quad := \Big(\min_{\pi} \Big((\|X\|_1 + \|Y\|_1 - 2\|\pi\|_1) C^p \\
&\qquad + \sum_{i,j,k,l} \rho_{(i,j),(k,l)} \pi_{(i,j),(k,l)} \Big) \Big)^{1/p},
\end{aligned}
\tag{17}
$$

where the minimum is taken over all $\pi = (\pi_{(i,j),(k,l)}) \in \mathbb{R}_+^{r \times s} \times \mathbb{R}_+^{r \times s}$ such that

$$
\sum_{k,l} \pi_{(i,j),(k,l)} \leq x_{i,j},
$$

$$
\sum_{i,j} \pi_{(i,j),(k,l)} \leq y_{k,l}
$$

for all $i, k \in [r]$, $j, l \in [s]$ and, furthermore,

$$
\|X\|_1 = \sum_{i,j} x_{i,j}, \quad \|Y\|_1 = \sum_{k,l} y_{k,l},
$$

$$
\|\pi\|_1 = \sum_{i,j,k,l} \pi_{(i,j),(k,l)}.
$$

In this way, $\pi_{(i,j),(k,l)}$ may be interpreted as the amount of mass transported from pixel (i,j) to pixel (k,l). Clearly, d_{UBW} is in the same spirit as the TT metric in the previous subsection, and indeed both definitions derive the same unbalanced Wasserstein metric on more abstract spaces of measures. The metric d_{UBW} and the computation of corresponding Fréchet means was recently studied in [26]. Note that the parameter C has a slightly different scaling in that paper. We compute d_{UBW}-distances and their barycenters on the given pixel grid with the R packages transport [39] and WSGeometry [25], respectively. The latter provides an adapted version of the *Matrix-based Adaptive Alternating Interior-Point Method (MAAIPM)* by [19] for this task.

6. SIMULATION STUDIES

In what follows, we investigate the different statistics from Table 1 for simulated point pattern and image data under various distributions.

In spatial statistics, there are usually two fundamentally different ways how point process distributions can deviate from CSR. One is spatial inhomogeneity of points, that is, points may be more or less likely to occur in different regions of the space. The ability of tests to detect deviations from CSR against various spatially inhomogeneous alternatives is studied in Section 6.1. The other way is interaction of points, that is, presence of points in one region may excite or inhibit the presence of other points nearby. In Section 6.2, we study how well the statistics discern between various interaction strengths in homogeneous Strauss processes. We consider simulated image data in Section 6.3.

For the evaluations in Sections 6.1 to 6.3, we perform permutation tests. These are based on generating M independent uniform permutations of the indices of the group elements resulting in alternative split-ups of the data into k groups of sizes n_i, $1 \leq i \leq k$. We then determine the rank r of the statistic-value for the original split-up within the statistic-values of the alternative split up (from $r = 1$ for the highest value to $r = M + 1$ for the lowest value).

It is easily checked (and well known) that $p = \frac{r}{M+1}$ is an honest p-value (i.e., $\mathbb{P}(p \leq \alpha) \leq \alpha$ for every $\alpha \in (0, 1)$). We reject the null if $p \leq 0.05$.

In all of these tests, we use $M = 999$ permutations if no barycenter computation is needed and $M = 99$ permutations if barycenter computation is needed. In view of the enormous number of possible split-ups ($\binom{40}{20} \approx 1.4 \cdot 10^{11}$ in most of the experiments considered below), this means that there is a high degree of randomization in each individual test. The small $M = 99$ was necessary due to the large computational burden of computing pseudo-barycenters in point pattern space (see Section 5). For statistics that do not require barycenter computation, choosing $M = 999$ typically results in much faster computation time than the choice of $M = 99$ for statistics that do require barycenter computation. For reproducibility of individual test results, a higher M or (where possible) comparing within all possible split-ups into groups would be desirable in both cases.

Preferring exact permutation tests over tests based on the limit χ^2-distribution is in agreement with the recommendations from previous papers and corresponds to our own experience. However, the χ^2-approximation of our L-statistic is quite fast as we can see in Appendix B, where we compare the finite sample distributions of the new L- and the Dubey–Müller statistics.

In all tests, we use as the underlying space $\mathcal{R} = [0, 1]^2 \subset \mathbb{R}^2$ with the Euclidean metric. The significance level is always $\alpha = 0.05$. Furthermore, we choose as order $p = 2$ and as penalty $C = 0.25$, which means that $\sqrt{2} \cdot 0.25 \approx 0.35$ is the maximal contribution that a single matched point pair makes to the corresponding TT-distance (or a single unit of mass transported makes to the unbalanced Wasserstein distance in the image case), that is, the actual Euclidean distances are cut off at this value. In applications, the choice of C is often based on the physical reality of the data and possibly the goal of the analysis. For the present simulation study, we tried not to restrict a substantial proportion of matching distances while keeping the contribution of additional points or additional pixel mass reasonably low. In the point pattern case, Table 2 gives an overview of the ratio of point pairs matched above the cutoff distance relative to pairs matched below the cutoff distance for various values of C based on pairwise comparisons of 1000 point patterns simulated according to CSR with intensity $\lambda = 35$. For $C = 0.25$ we have for every matching above the cutoff distance $1/0.038 \approx 26$ matchings below the cutoff distance.

TABLE 2

Pairwise comparison within 1000 *patterns simulated from CSR on* $[0, 1]^2$ *with intensity* $\lambda = 35$ *for various penalties* C. *The first two columns give the ratios of the point pairs above cutoff and the unpaired points, both relative to the point pairs below cutoff. The last column is the mean* d_{TT}*-distance among the* $\binom{1000}{2}$ *pairwise comparisons*

	Above cutoff	Unpaired	Mean d_{TT}
$C = 0.1$	0.424	0.311	0.309
$C = 0.15$	0.167	0.257	0.393
$C = 0.2$	0.078	0.24	0.457
$C = 0.25$	0.038	0.233	0.512
$C = 0.3$	0.017	0.23	0.561
$C = 0.35$	0.006	0.229	0.609

6.1 Inhomogeneity in Point Pattern Data

We compare $k = 2$ groups of $\tilde{n} = n_1 = n_2 = 20$ point patterns. Patterns in Group 2 are simulated from CSR with $\lambda = 35$. In Group 1, they are simulated from various inhomogeneous scenarios, that is, from Poisson process distributions where the intensity function (the density of the measure ν with respect to Lebesgue measure) deviates more or less from a constant but still integrates up to 35 over the whole window $\mathcal{R} = [0, 1]^2$.

In Scenarios 1–3, the intensity is obtained by adding a number of rotation-invariant Gaussian distributions with different means but the same covariance matrix $\sigma^2 I$ and scaling to total mass 35. For simplicity, we do not restrict the intensity to $\mathcal{R}$, but as can be seen from Figure 1 only very few points outside $\mathcal{R}$ occur. Scenarios 4–6 use as intensity an exponential function that is constant in the y-coordinate and induces a certain tendency for points to lie in the left part of the window rather than in the right part.

Table 3 provides more information about the chosen parameters. Figure 1 shows five example point patterns for each scenario. In addition, we add a Scenario 0, which corresponds to simulating the first group also from CSR with $\lambda = 35$.

TABLE 3

Overview of the Poisson process intensities for the six scenarios. The proportionality constant is chosen such that the expected number of points in each scenario is 35. *By* φ_{μ,σ^2}, *we denote the density of the bivariate normal distribution with mean* $\mu \in \mathbb{R}^2$ *and covariance matrix* $\sigma^2 I$. *The different* μ_i *used are seen in Figure* 1

Scenario	$\lambda(x, y)$ proportional to
1	$\sum_{i=1}^{3} \varphi_{\mu_i,0.075}(x, y)$
2	$\sum_{i=1}^{3} \varphi_{\mu_i,0.1}(x, y)$
3	$\sum_{i=1}^{4} \varphi_{\mu_i,0.1}(x, y)$
4	$\exp(-2x)$
5	$\exp(-1x)$
6	$\exp(-0.02x)$

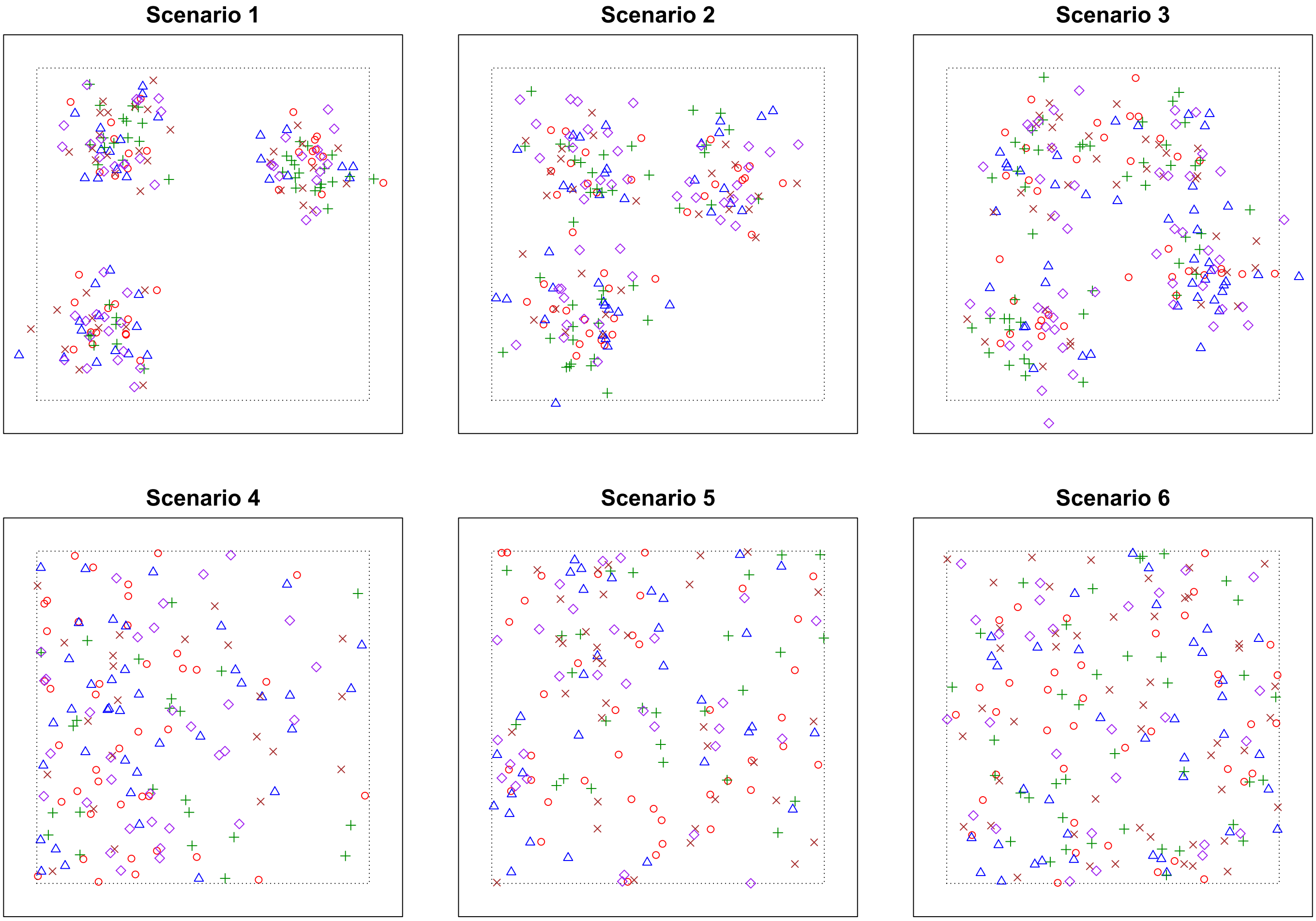

FIG. 1. *Sample patterns for the six scenarios in Section* 6.1. *For each scenario there are five patterns simulated from the same inhomogeneous Poisson process distribution, which are depicted with individual colors and symbols. The intensities* (*up to constants*) *are given in Table* 3. *The dotted line delimits the window* $[0, 1]^2$.

Table 4 gives the results in terms of numbers of rejections (out of 100) of the null hypothesis of equal distribution in both groups.

We observe that the direct ANOVA procedures perform much better than the Levene (or indirect ANOVA) procedures. This is not so surprising, because the inhomogeneity experiment considers two groups of distributions that are different in terms of their location in the point pattern space. To see this intuitively, think about the distributions in Scenarios 1–6 (and 0 as a boundary case) in terms of producing locally perturbed versions of a typical point pattern, which is more or less any one of the example point patterns in Figure 1 (more appropriately one would rather think of an idealized version of these patterns, such as the Fréchet mean). Among the direct ANOVA methods, Anderson F_A performs substantially better than Fréchet T_F and has still a reasonable chance to detect the faint differences between Scenarios 6 and 0 when presented with the 20 patterns from each group. Our new L-test performs somewhat better than the Fréchet L-test, but both tests are only able to detect the inhomogeneity (with reasonable probability) when it is very obvious (Scenarios 1–3).

TABLE 4
Numbers of rejections of the null hypothesis “equal distribution in both groups” based on 100 *data sets per column. In each data set, the first group is sampled from the scenario indicated in the column and the second group is sampled from Scenario* 0

Scenario	1	2	3	4	5	6	0
Anderson F_A	100	100	100	100	99	39	2
New L	93	76	77	14	7	9	3
Fréchet T_F	100	100	100	99	11	0	4
Fréchet T_L	59	24	47	13	9	12	4

To give an impression of the unbalanced situation, we repeat the experiment with groups of sizes $n_1 = 10$ (Scenarios 0–6) and $n_2 = 30$ (Scenario 0); see Table 5.

As one might have expected, the empirical power deteriorates compared to the balanced setting, but the effect is not particularly strong. The relative performance of the various tests remains largely the same as in the balanced setting.

TABLE 5

Numbers of rejections of the null hypothesis "equal distribution in both groups" based on 100 *data sets per column in the unbalanced setting. In each data set, the first group is sampled from the scenario indicated in the column and the second group is sampled from Scenario* 0

Scenario	1	2	3	4	5	6	0
Anderson F_A	100	100	100	100	90	33	5
New L	81	53	55	9	4	12	8
Fréchet T_F	100	100	100	100	76	29	7
Fréchet T_L	51	27	36	11	3	6	4

6.2 Interaction in Point Pattern Data

Again we compare groups of $\tilde{n} = 20$ point patterns. This time the group distributions are homogeneous (stationary) but differ in the degree of point interaction. For this, we consider the distribution of the homogeneous Strauss process on the unit square $\mathcal{R} = [0, 1]^2$, which is obtained by specifying the density $f : \mathfrak{N}_{\text{fin}} \to \mathbb{R}_+$,

$$f(\xi) := c \cdot \beta^{|\xi|} \cdot \gamma^{s_R(\xi)},$$

with respect to CSR with intensity 1 on $\mathcal{R}$, where

$$s_R(\xi) = \sum_{\{x,y\} \subset \xi} \mathbb{1}\{\|x - y\| \leq R\}$$

is the number of pairs of points at distance $\leq R$ from one another. Here, $R > 0$ is the range of the interaction, $\beta > 0$ is the so-called activity (which controls the intensity of the process via an increasing function, that is however only accessible numerically) and $\gamma \in [0, 1]$ is the strength of the interaction. The constant c normalizes the density to an overall integral of 1 and is also not available in closed form. We write Strauss$(\beta, \gamma; R)$ for this point process distribution. Intuitively, a Strauss$(\beta, \gamma; R)$ process is obtained from a CSR(β) process by penalizing each outcome according to a factor γ per R-close point pair. Correspondingly, we have Strauss$(\beta, 1; R) =$ CSR(β) (regardless of R). At the other end of the spectrum, Strauss$(\beta, 0; R)$ is the distribution of a hard core process with no points allowed within distance R of other points.

For the simulation, we set $R = 0.1$ and consider scenarios based on the six different values $\gamma = 0$, 0.2, 0.4, 0.6, 0.8, 1. The activity β is adapted so that each time $\lambda = 35$. Figure 2 shows one realization for each of the six scenarios.

A comparison of all six groups in a single ANOVA yields perfect rejection for almost all test statistics (T_F only rejects 95 of 100). To have a reasonable basis for comparing the statistics, we perform separate omnibus tests for the three groups with $\gamma \in \{0, 0.2, 0.4\}$ (referred to as small γ), and the three groups with $\gamma \in \{0.6, 0.8, 1.0\}$ (large γ). Subsequently, we also test between each pair of groups within the small and the large setting. Table 6 gives the results in terms of numbers of rejections out of 100.

In contrast to the situation in the previous subsection (different inhomogeneity), we now observe that the *indirect* ANOVA procedures, that is, the Levene-type tests, perform considerably better than the direct ANOVA procedures. Again this is intuitively understandable because a small γ in the Strauss process leads to less dispersion, both in terms of a smaller variance for the total number of points and also with respect to typical distances of points from one another: for small γ, the points are quite regularly placed, whereas for larger γ there are erratic patches that are free of points leading typically to some points that have to be matched over longer distances, which in the squared Euclidean metric has quite some influence. A small γ will also lead to smaller average distances than a larger γ (either between point patterns or relative to a barycenter), which may explain why the difference in the performance of the indirect and direct ANOVA tests is somewhat less pronounced than in the inhomogeneity experiment.

Note again that the powers of the tests based on pairwise distances are slightly better than those of the tests based on barycenters.

6.3 Image Data

In our final simulation study, we consider image data, using the concepts and notation from Section 5.2. We always compare 2 groups of $\tilde{n} = n_1 = n_2 = 10$ images of size 16 × 16. Compared to real images from biomedicine, remote sensing or similar disciplines this is a mere toy example, but the fact that we want to include Fréchet ANOVA with exact barycenters and perform 100 tests per experiment (albeit again with only 99 permutations) essentially restricts us to such a data size. If it were for a single Anderson ANOVA and new Levene test based on a moderate number of group elements, images of sizes up to 128 × 128 would be unproblematic, and sizes beyond that would be well possible depending on sparsity features of the images and computation power invested. In all cases (including Fréchet ANOVA), one could also resort to dimensionality reduction techniques ranging from simple coarsening of larger images to the use of regularized optimal transport methods if very fine features of the original images are deemed not so important.

In our simulation scenarios, we consider "abstract" images obtained by simulating from random fields with values in $[0, 1]$ to represent grayscale pixel values. For the pixels, we use a regular subdivision of the unit square $[0, 1]^2$, placing pixel centers at $(u, v) \in G := \{\frac{1}{32}, \frac{3}{32}, \ldots, \frac{31}{32}\}^2$. Then we obtain images as $(f(Z_{(u,v)}))$, where $f : \mathbb{R} \to (0, 1)$, $f(x) = (1 + \exp(-x))^{-1}$, is the standard logistic function and $(Z_{(u,v)})$ is a Gaussian random field with mean zero and exponential covariance

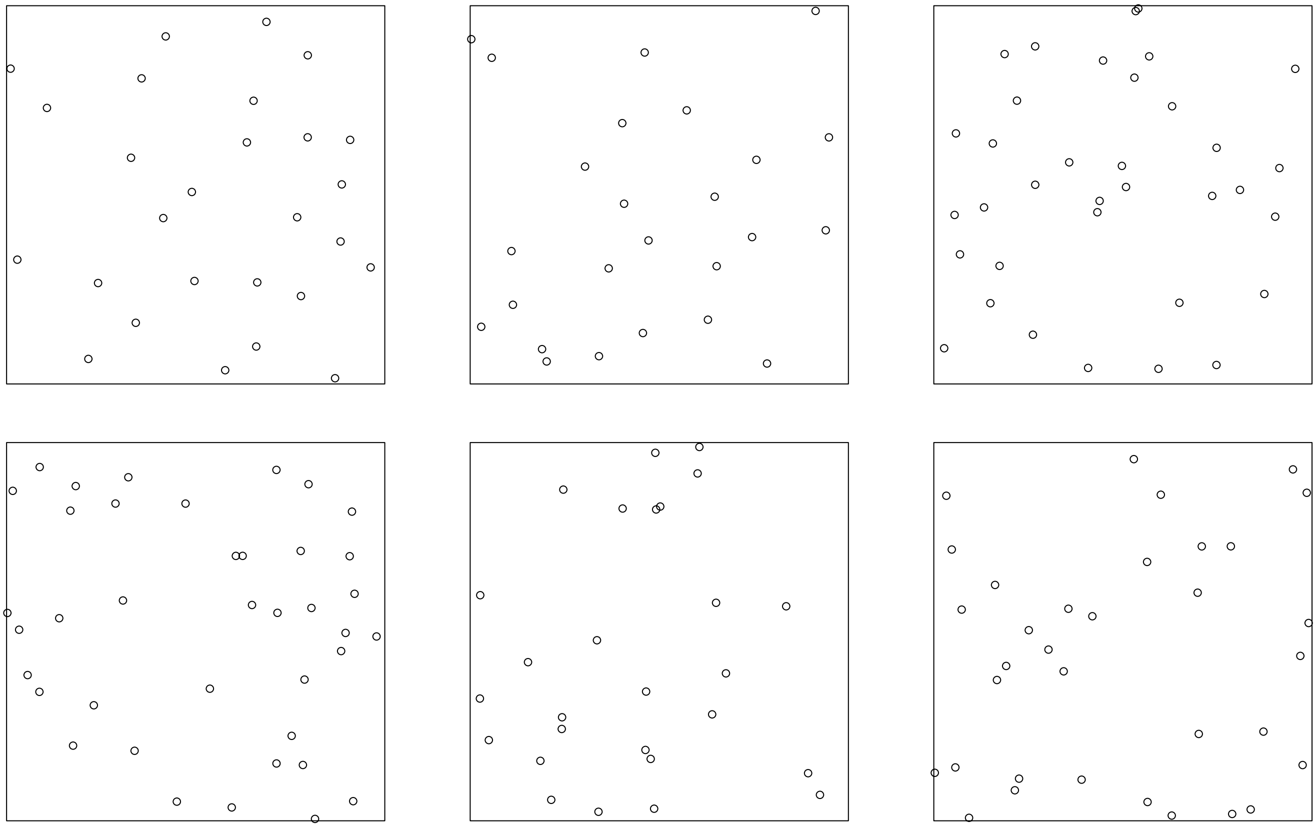

FIG. 2. *Simulations from* Strauss$(\beta, \gamma; 0.1)$*-distributions, where rowwise from left to right* $\gamma = 0, 0.2, 0.4, 0.6, 0.8, 1$ *and* β *is adjusted such that* $\lambda = 35$. *For* $\gamma = 0$ *we have a realization of a hard core process, for* $\gamma = 1$ *a realization from CSR.*

function, that is, $(Z_{(u,v)})_{(u,v)\in G}$ is multivariate Gaussian with $\mathbb{E}Z_{(u,v)} = 0$ and

$$\operatorname{Cov}(Z_{(u,v)}, Z_{(u',v')}) = \sigma^2 \exp\left(-\left\|\begin{pmatrix}u\\v\end{pmatrix} - \begin{pmatrix}u'\\v'\end{pmatrix}\right\|_2 / \gamma\right)$$

for some variance $\sigma^2 > 0$ and a length scale parameter $\gamma > 0$. We try again to capture appropriate concepts for different locations and different dispersions between the groups in our two experiments.

For the first experiment, we simulate two images as Gaussian random field with mean zero, standard deviation $\sigma = 1/6$ and length scale $\gamma = 0.1$. Then we add 10 times i.i.d. Gaussian noise with mean 0 and standard deviation $\sigma = 5/6$ to each image before applying the logistic function to arrive at the two groups. Figure 3 shows in the first column the underlying initial images for the two groups. These remain the same for all 100 repetitions of the experiment. Then there are four sample images for each group (rows) belonging to one particular repetition. The sample images all use the same color scale, whereas the scale of the initial images is exaggerated for better visibility.

For the second experiment, we directly use independent simulations from a Gaussian random field with mean zero and standard deviation $\sigma = 1$, but having different length scales of $\gamma = 0.1$ and $\gamma = 0.25$ for the two groups. Figure 4 gives five sample images for each group (rows) all on the same color scale. While in the first experiment the true difference between the groups is due to their varying

TABLE 6
Numbers of rejections of the null hypothesis "equal distribution in all groups" based on 100 *data sets per column. In each data set, the point patterns in all groups are sampled from a Strauss distribution with* $\lambda = 35$ *and* $R = 0.1$, *but different* γ. *The left-hand side of the table tests among groups sampled with small* γ, *the right-hand side among large* γ

$\gamma \in \{0, 0.2, 0.4\}$	Omnibus	0 vs. 0.4	0 vs. 0.2	0.2 vs. 0.4	$\gamma \in \{0.6, 0.8, 1\}$	Omnibus	0.6 vs. 1	0.6 vs. 0.8	0.8 vs. 1
Anderson F_A	87	90	55	17	Anderson F_A	6	8	8	4
New L	94	96	60	28	New L	59	67	18	20
Fréchet T_F	65	76	45	18	Fréchet T_F	6	8	7	6
Fréchet T_L	70	82	33	21	Fréchet T_L	46	53	16	17

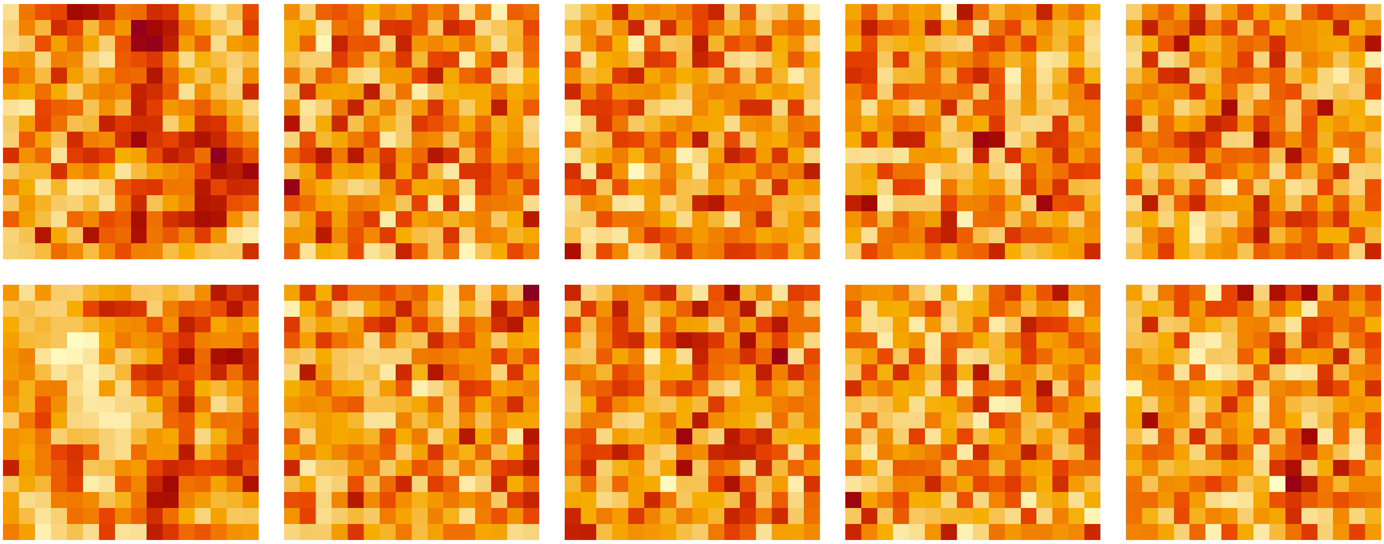

FIG. 3. *First image experiment. The first image in each row is the underlying image used for the construction of the data* (*color scale exaggerated*). *Then for each group* (*row*) *four sample images, obtained by adding i.i.d. noise, are shown, all using the same scale of inverse heat colors from* 0 (*light yellow*) *to* 1 (*dark red*).

average pixel intensity according to location, the group difference in the second experiment is based on the more global feature of correlation.

Table 7 lists the number of rejected tests out of 100 based on the unbalanced Wasserstein metric with parameters $p = 2$ and $C = 0.25$. We observe a similar phenomenon as for the point pattern data. In the first experiment, where the group data scatters around different means, the location tests based on statistics F_A and T_F perform considerably better. In the second experiment, where the groups scatter around the same mean, but to a different extent in terms of their average pixel value over the whole space, the dispersion tests based on the new L statistic and T_L perform much better. Again it appears that in addition to being several orders of magnitude faster than the Fréchet ANOVAs, the pairwise distance based tests also have the edge regarding discriminatory power.

7. APPLICATIONS

In this section, we apply the pairwise distance tests to real data examples. We investigate the location of bubbles in a mineral flotation experiment and count-based images of pests in the city of Madrid. The structure of the data calls for a two factor design. For both the Anderson ANOVA and our new L, the corresponding statis-

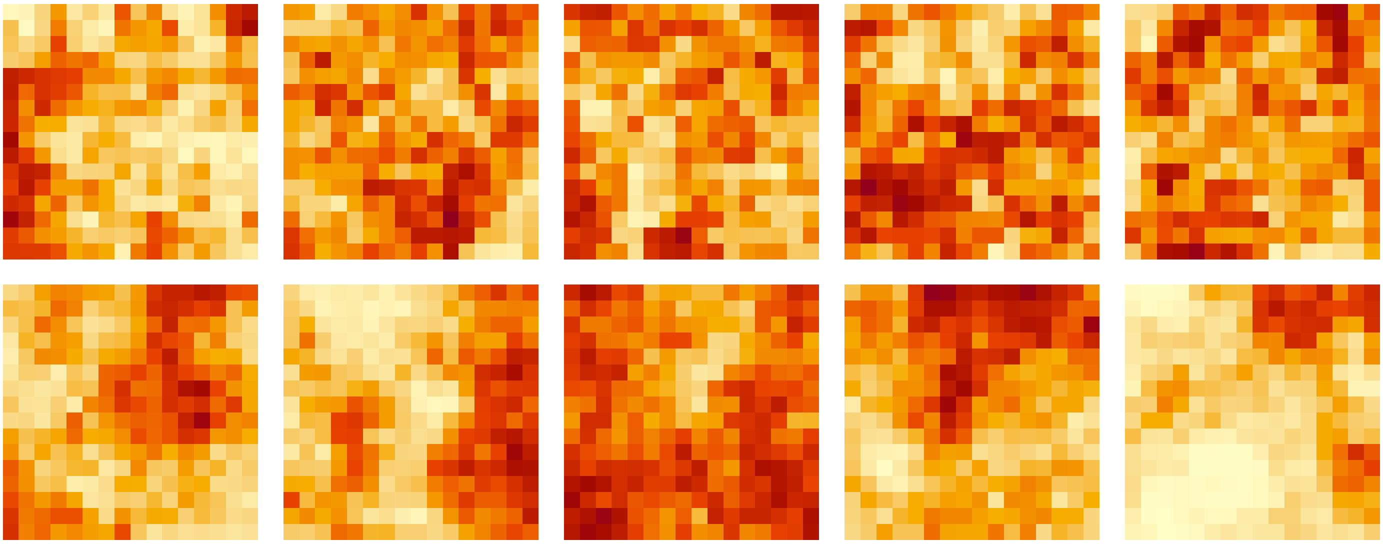

FIG. 4. *Second image experiment. Five sample images for each group* (*row*) *are shown, all using the same scale of inverse heat colors from* 0 (*light yellow*) *to* 1 (*dark red*). *The samples have been generated from logit-Gaussian random fields with length scales* $\gamma = 0.1$ *in the top row and* $\gamma = 0.25$ *in the bottom row.*

TABLE 7

Numbers of rejections of the null hypothesis "equal distribution in all groups" based on 100 *data sets per column. In the first experiment, images in the two groups have been generated by adding Gaussian white noise to two different initial images. In the second experiment, images in the two groups have been generated from Gaussian random field distributions differing in the length scales* $\gamma = 0.1$ *and* $\gamma = 0.25$ *of their* (*exponential*) *covariance function*

	First experiment	Second experiment
Anderson F_A	45	26
new L	4	78
Fréchet T_F	39	14
Fréchet T_L	4	65

tics are straightforward to derive from the classic balanced two-way ANOVA (see Section 2). For Anderson ANOVA, these statistics are given in [2], for our new L we provide them below. We do not consider the Fréchet statistics in this section because their two-way formulation is not so straightforward, and we did not find a recommended version, neither in [16] nor elsewhere.

7.1 Balanced Two-Way Levene's Test

As mentioned in Section 4.1, it is easy to generalize statistic (5) to a two-way design, which will further be useful for the two applications analyzed in this section.

Suppose we have independent observations $x_{i_1 i_2 j} \in \mathcal{X}$, $1 \leq j \leq \tilde{n}$, $1 \leq i_1 \leq k_1$, $1 \leq i_2 \leq k_2$ from groups obtained by crossing a Factor a with k_1 levels and a Factor b with k_2 levels with $\tilde{n}$ observations for each combination. In a similar way as above, we denote by $d_{i_1 i_2 j}$ the jth half-distance in the group (i_1, i_2), where $j = 1, \ldots, \tilde{N} := \binom{\tilde{n}}{2}$. Set then

$$\mathrm{RSS} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{\tilde{N}} (d_{i_1 i_2 j} - \bar{d}_{i_1 i_2 \cdot})^2,$$

$$\mathrm{MSS} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \tilde{n}(\bar{d}_{i_1 i_2 \cdot} - \bar{d}_{\cdots})^2,$$

$$\mathrm{SSa} = \sum_{i_1=1}^{k_1} k_2 \tilde{n}(\bar{d}_{i_1 \cdot\cdot} - \bar{d}_{\cdots})^2,$$

$$\mathrm{SSb} = \sum_{i_2=1}^{k_2} k_1 \tilde{n}(\bar{d}_{\cdot i_2 \cdot} - \bar{d}_{\cdots})^2,$$

$$\mathrm{SSi} = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \tilde{n}(\bar{d}_{i_1 i_2 \cdot} - \bar{d}_{i_1 \cdot\cdot} - \bar{d}_{\cdot i_2 \cdot} + \bar{d}_{\cdots})^2,$$

where the various means are taken over the dot components in the usual way. Note that we never use any distances between observations of different factor combinations.

In addition to the omnibus test for group differences as in one-way ANOVA, we may then perform Levene-type tests for effects of Factors a and b separately, as well as for an interaction effect. The corresponding statistics are

$$L = \frac{N - k_1 k_2}{(k_1 k_2 - 1)} \frac{\mathrm{MSS}}{\mathrm{RSS}}, \quad La = \frac{N - k_1 k_2}{k_1 - 1} \frac{\mathrm{SSa}}{\mathrm{RSS}},$$

$$Lb = \frac{N - k_1 k_2}{k_2 - 1} \frac{\mathrm{SSb}}{\mathrm{RSS}}, \quad Li = \frac{N - k_1 k_2}{(k_1 - 1)(k_2 - 1)} \frac{\mathrm{SSi}}{\mathrm{RSS}}.$$

7.2 Bubbles in a Mineral Flotation Experiment

We consider the data from [21], which provides locations of bubbles in a mineral flotation experiment, where the interest is analyzing if the spatial distribution might be affected by frother concentrations and volumetric airflow rates. Indeed, the data set consists of 54 images containing a total of 8385 floating bubbles. The images of bubbles can be regarded as spatial point patterns where the centroids of the bubbles correspond to the points. In addition, we have three frother concentration levels (5 ppm, 10 ppm, 15 ppm) as well as three volumetric airflow rate levels (5 l/min, 8 l/min, 10 l/min), and we have six replicates of point patterns at each combination of levels of such factors. The treatment combinations of the experiment, as well as the observed bubble point patterns, are represented in Figure 5.

We used the two-way design of Levene's statistic from Section 7.1 to test for influence of the individual factors and the interaction, as well as for an overall difference between the groups in the context of the point pattern space introduced in 5.1. For comparison, we also used the two factor statistics from [2], we performed a two-factor ANOVA on the number of points per pattern, and finally complemented our analysis with a two-factor ANOVA with K-functions, so as to link our analysis with that of [21]. We did a permutation test with 999 permutations.

In Section 6, the cutoff was always fixed to $C = 0.25$. This was a reasonable value for point patterns with expected 35 points in the unit square. In the bubble data, the number of points per observed pattern ranges from 21 to 353. With such a great variability in the number of points, we suggest adjusting the cutoff to prevent that distances between two patterns are dominated by their different numbers of points. For the results presented in this section, we computed the mean number of points of the tested patterns $\bar{n}$ and used the cutoff $\bar{C} = 0.25 \cdot 35/\bar{n}$ for the computations of d_{TT}. For more details to the cutoff, see (16).

The p-values of the permutation tests are shown in Tables 8 and 9. In particular, Table 8 shows results for the whole data set, while Table 9 depicts results for only part of the data, leaving out the third column, that is, any patterns from frother concentration of 15 ppm. In both cases, our new Levene, Anderson F_A, the ANOVA on the number of points per pattern, and the ANOVA for K-functions

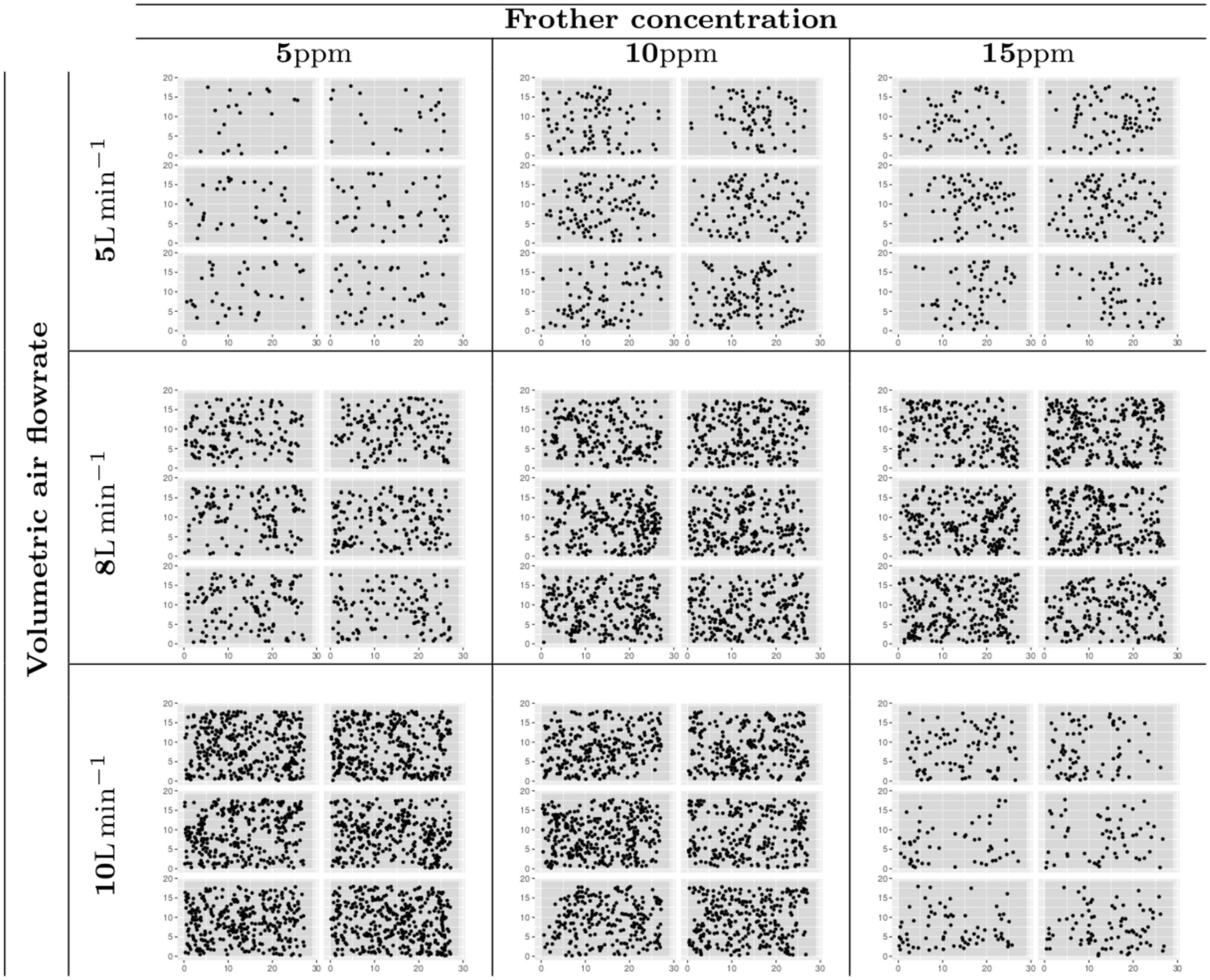

FIG. 5. *Arrangement of floating bubbles data. Rows represent the three frother concentration levels and columns the three volumetric air flowrate levels* (*treatments*). *Each cell contains six spatial point patterns* (*responses*).

detect significant influence of each of the two factors and the interaction. We already recommended to always perform both, the tests for differences in variability and the test for differences of means. In the second test scenario, both Levene's test and Anderson F_A detect significance for the frother concentration and the interaction of both parameters for our usual significance level of 5%. But the

TABLE 8
Results of the different tests for the bubble data. Quantiles are obtained by a permutation test with 999 *permutations. The cutoff is* $C = 0.0564$, *the maximal radius for the* K*-functions is* $r = 0.15$

p-values	FC	VA	Interaction	Overall
Anderson F_A	0.003	0.001	0.001	0.001
new L	0.001	0.001	0.001	0.001
Number of points	0.001	0.001	0.001	0.001
K-functions	0.005	0.001	0.001	0.001**

** This is the p-value for the sum of both factors, not the overall ANOVA statistic.

TABLE 9
Results for the different tests for the bubble data, leaving out the frother concentration of 15*ppm. Quantiles are obtained by a permutation test with* 999 *permutations. The cutoff is* $C = 0.0636$, *the maximal radius for the* K*-functions is* $r = 0.15$

p-values	FC	VA	Interaction	Overall
Anderson F_A	0.043	0.001	0.019	0.001
New L	0.001	0.001	0.001	0.001
Number of points	0.001	0.002	0.001	0.001
K-functions	0.002	0.022	0.002	0.006**

** This is the p-value for the sum of both factors, not the overall ANOVA statistic.

relative difference between the p-values of the two tests is very large. For the smaller significance level of 1%, our Levene's test still detects significance where Anderson F_A does not. So, the test for differences of means might not be enough in a practical application. This is particularly important in cases where, as it is the case for the bubble data, the number of points plays a crucial role in the behavior and structure of the point patterns.

We see that for this data apparently the numbers of points per pattern contain enough information to detect significant influence of the factors. This is not very surprising since the number of points per pattern is similar in the 6 patterns of a single cell, but the differences between cells are large.

This observation is reinforced by a classical multidimensional scaling (mds). Based on the TT-distances between the point patterns, we translated every point pattern into a single point in $\mathbb{R}^2$. The mds was applied first for the whole bubble data set (see Figure 6) and then for a subset of the data consisting of the first and second columns, leaving out the data with a frother concentration of 15 ppm; see Figure 7. This is the same data that we used for our analyses in Tables 8 and 9. The three levels of the air flow are encoded by the colors "red," "green" and "blue," same color means same air flow rate, and the three levels of the frother concentration are encoded by the symbols "circle," "triangle" and "cross." When we compare these plots to the images of the point patterns in Figure 5, we can see that the multidimensional scaling sorts the point patterns from left to right in ascending order by their number of points per pattern. In Figure 7, we can see that the points that correspond to the data with a frother concentration of 5 ppm, that is, the circles and the data with a frother concentration of 10 ppm, that is, the triangles are scattered differently. The (coordinatewise) means of the triangles and circles are similar, but we can see that the circles are more scattered along both axes. We conjecture that it is this difference in scatter that our Levene's test is able to detect in Table 9, whereas the Anderson F_A only barely detects a slight difference in means.

7.3 Pests in the City of Madrid

As a second application, we look at the seasonal distribution of rats and cockroaches in the city of Madrid over the years 2010–2013. The data is based on reports by citizens of either direct sightings or clear traces of the presence of these pests. A small subset of this data, focusing on the rats in a single district, was analyzed in [41] on a finer time and space resolution and with further covariates.

For the present analysis, we consider spatial histograms based on count data within contiguous bins of 500 × $500m^2$. Figure 8 shows the seasonal counts for the year 2013 as an example. The total counts over the years have been fairly stable at 3000 to 4000 reported sightings without a clear trend. We therefore use the year for replication and perform a 2-way comparison with the factors *season* and *species* in the setting of the image space introduced in Subection 6.3. As parameters for the unbalanced Wasserstein metric, we choose $p = 2$ and $C = 2000m$. The latter choice seems like a good upper bound on the interaction radius of rats and cockroaches over longer periods of time, but the results are quite robust to other choices, such as $C = 5000m$.

A first comparison is given in Table 10. We can understand from Figure 8 (and similar observations for the other 3 years) why almost all the tests are highly significant. There is a clearly visible seasonal trend, which is particularly pronounced for the cockroaches. One possible explanation for the interaction term is that the sea-

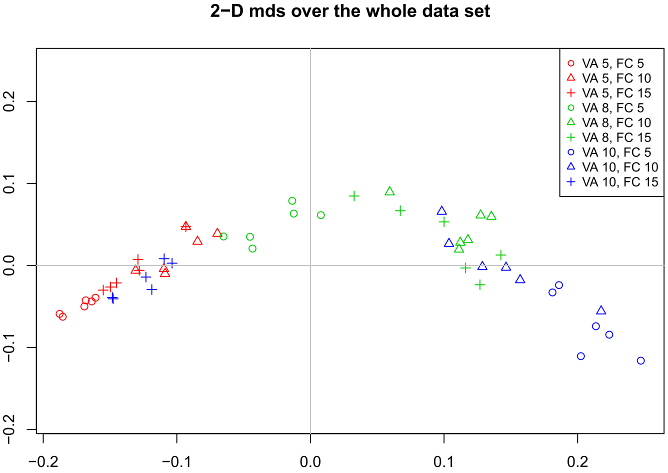

FIG. 6. *The bubble data after a multidimensional scaling into two dimensions based on the distance matrix w.r.t. the TT-metric. Colors according to volumetric airflow rate* (*VA*); *symbols according to frother concentration* (*FC*).

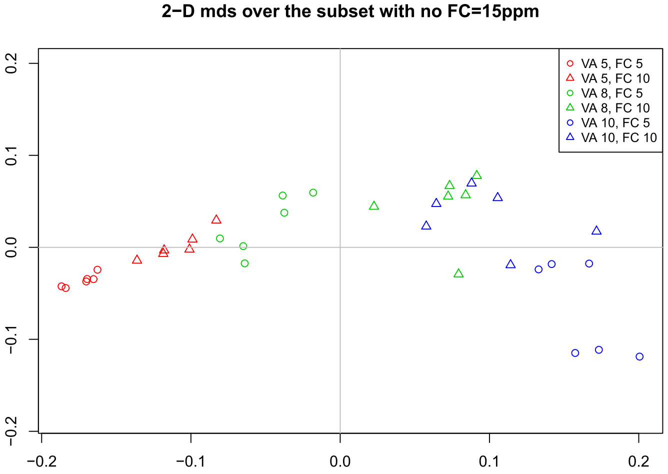

FIG. 7. *The bubble data without a frother concentration of* 15 *ppm, after a multidimensional scaling into two dimensions based on the distance matrix w.r.t. the TT-metric. Colors according to volumetric airflow rate* (*VA*); *symbols according to frother concentration* (*FC*).

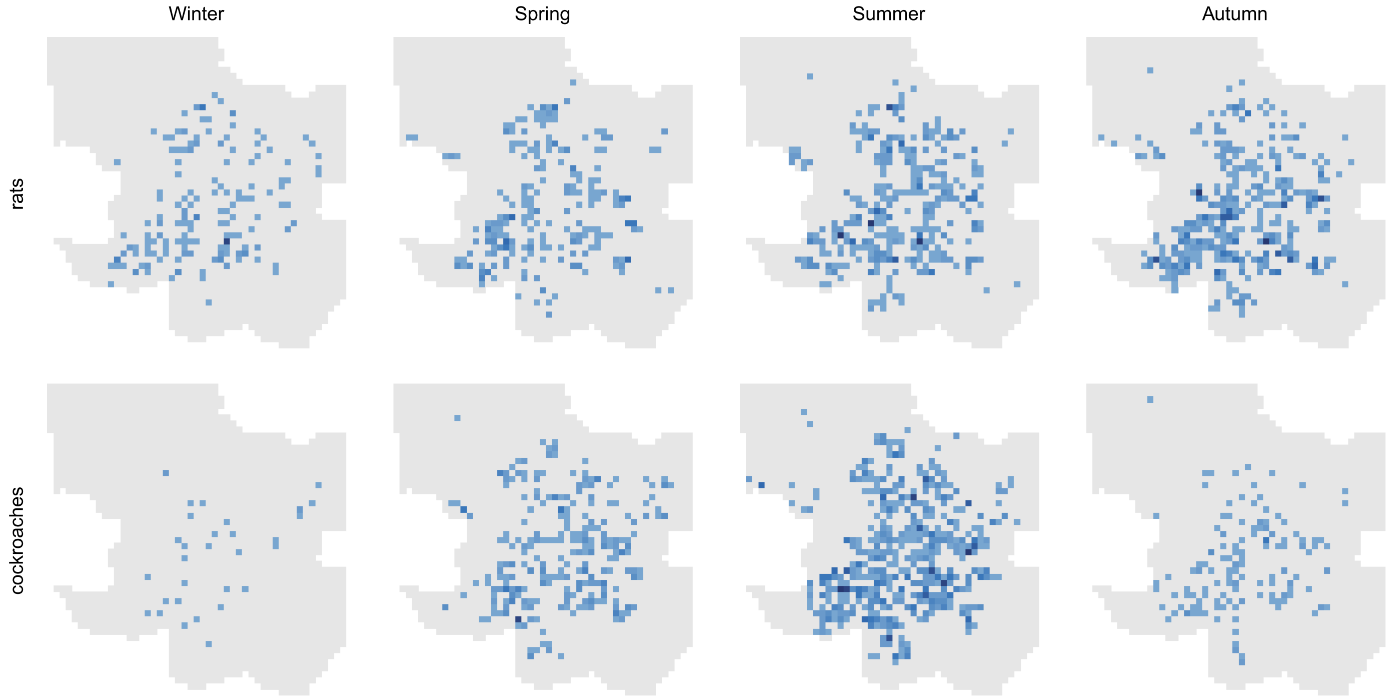

FIG. 8. *Seasonal pest counts in* 2013 *in the city of Madrid. Numbers of sightings of rats and cockroaches* (*or their traces*) *counted in bins of* 500 × $500m^2$. *Colors range from light blue* (1 *sighting per season*) *to dark blue* (10 *or more sightings per season*) *with a maximum of* 16 *being reached in summer for both rats and cockroaches.*

sonal cycles are shifted: while the cockroaches reach their highest spread in spring and especially summer and face a substantial decline in autumn, the highest numbers of rats are observed somewhat later in the year, with about equal numbers in summer and autumn. The new L statistic does not pick up a significant difference between the species, which might be hidden behind the strong interaction effect: if we perform ANOVA only for the species the p-value for the L statistic is 0.007.

In a similar way as in the previous example, it is to be expected that several of the differences we observe are substantially influenced by the total counts in the images, which differ especially during the winter months, see Table 11. In a second analysis it is therefore natural to ask for differences in the *spatial distribution* of the pests, regardless of their total counts. For this, we normalize each image so that its pixel values sum up to one. Thus, each image represents the ratio of sightings per bin for the respective species and season. We may then use the usual (balanced) Wasserstein-2 metric for comparison, which means in terms of equation (17) that we remove the first summand, which has the factor C^p in it. Table 12 shows the resulting p-values.

So, even without the influence of the total counts the results remain largely the same, indicating differences in the spatial distributions according to species and season (and their interaction). Again the new L test for species is not significant, but a new L test for species alone would be highly significant with a p-value of 0.001.

We have opted here for seasonal rather than the monthly data to make it easier to present and discuss the data. For completeness, we mention that in a full 2-way ANOVA with monthly data *all* the tests (including the new L) give a p-value of 0.001.

As a series of follow-up test, we compare the spatial distributions of the species during their high seasons, that

TABLE 10
Results of the different tests for the Madrid pest data based on the unbalanced Wasserstein-2 metric between images. Quantiles are obtained by a permutation test with 999 *permutations and the cutoff is* $C = 2000m$

p-values	Species	Season	Interaction	Overall
Anderson F_A	0.001	0.001	0.001	0.001
New L	0.38	0.001	0.001	0.001

TABLE 11
Total numbers of pest sighting per year, species and season (*winter, spring, summer, autumn*)

	Rats				Cockroaches			
	Win	Spr	Sum	Aut	Win	Spr	Sum	Aut
2010	199	295	488	383	31	635	1309	183
2011	255	485	659	655	47	765	1064	190
2012	229	435	591	375	38	673	1293	123
2013	190	321	534	615	32	378	901	146

TABLE 12
*Results of the different tests for the Madrid pest data based on the usual Wasserstein-*2 *metric between normalized images. Quantiles are obtained by a permutation test with* 999 *permutations*

p-values	Species	Season	Interaction	Overall
Anderson F_A	0.001	0.001	0.007	0.001
New L	0.138	0.001	0.001	0.001

is, rats in summer (RS), rats in autumn (RA) and cockroaches in summer (CS). See Table 13 for the results. There are clear overall differences between these groups, which are due to different spatial distributions of rats and cockroaches. It is notable that the p-values between the groups "rats in summer" and "rats in autumn" are exactly one, which may be explainable by a strong correlation between the (normalized) images of subsequent seasons of the same year, whereas the repetitions within the groups, stemming from different years, are more independent.

8. CONCLUSIONS AND FURTHER DISCUSSION

In this paper, we have given an overview of four recent ANOVA-like procedures for data in general metric spaces, which according to their construction can be cross-classified as in Table 1.

Comparing the rows of this table, we conclude that although the tests based on Fréchet means have high theoretical appeal inasmuch as they precisely transfer the concepts of means and variances to a general metric space, they are often hampered by the high computational cost of such (approximate) means. They also tend to perform somewhat worse than their pairwise distance based counterparts in our examples. Comparing the columns of Table 1, we have discovered that testing for location and dispersion translates in our spatial settings into detection power of differing spatial inhomogeneity and differing spatial dependence, respectively. In practice, unless there is special interest in one of these phenomena (e.g., because the data generation process precludes the other one), we recommend to conduct both tests of a single row. These tests complement each other, so that in addition to test decisions we also obtain insight into the nature of the departure from the null hypothesis.

TABLE 13
Results of follow-up tests for the three groups "rats in summer" (*RS*), *"rats in autumn"* (*RA*) *and "cockroaches in summer"* (*CS*) *using the Wasserstein-*2 *metric between normalized images. Quantiles are obtained by a permutation test with* 999 *permutations*

p-values	3 groups	RS vs. RA	RS vs. CS	RA vs CS
Anderson F_A	0.009	1	0.029	0.031
New L	0.008	1	0.023	0.029

Regarding our own contribution to this table, the new L statistic, we were able to establish its utility in an important and rather general situation: If there are several groups of data differing in terms of their spatial dependence and the Fréchet means are expensive to compute, then the new L test is clearly the method of choice. It has much higher power than the location based tests and somewhat higher power than the T_L test. With respect to the latter, it is faster by several orders of magnitude, for example, the 100 permutation tests in Section 6.2 took only 2 seconds for the new L (999 permutations per test), but about 45 minutes for T_L (in spite of only 99 permutations per test).

In summary, we find that the new L in combination with the Anderson F_A has a slightly better performance and allows for considerably faster computation than the other methods in settings where the computation of barycenters is costly.

There are two points deserving further attention. On the one hand, we might want to combine location and dispersion tests in a single test decision. As always, this can be achieved by controlling the criterion seen as relevant (e.g., the familywise error rate) with an appropriate correction procedure. [16] suggest to consider a single test for the statistic $T = T_L + T_F$. Our experience from additional experiments shows that the performance of T lies between the performance of T_L and T_F, sometimes performing almost as good as the better of the two statistics in terms of power (e.g., for the balanced experiment in Section 6.1), but sometimes performing considerably worse (e.g., for the pairwise comparisons in Section 6.2).

Another point is the influence of the cutoff parameter C in the metrics, which determines at what cost mass can be created or destroyed and, thereby, up to what distance mass is transported. In Section 6, we chose $C = 0.25$ so as to keep the proportion of distances cut off small while not making the penalty for extra mass overly large. If instead we choose $C = 0.1$, say, we observe that the Anderson F_A and the Dubey–Müller T_F statistics considerably lose power to detect interaction differences (in Table 6 we would go from 55 to 11 rejections for F_A and from 45 to 14 rejections for T_F when testing $\gamma = 0$ vs. $\gamma = 0.2$), while our new L and the Dubey–Müller T_L considerably gain power (from 60 to 99 rejections for L and from 33 to 87 rejections for T_L when testing $\gamma = 0$ vs. $\gamma = 0.2$). This indicates that a judicious choice of C can make an important difference.

For future research, one might take a closer look at the estimator of the covariance γ used in our $\widetilde{L}$ statistic. This estimator is not unbiased, and it remains open if a statistic with an unbiased estimator works even better, in particular, for an asymptotic test.

Also it would be interesting to study the four procedures in Table 1 theoretically and empirically for 2-way ANOVA, both in the crossed balanced situation of

Section 7 and for more complex (possibly unbalanced) designs. Relevant applications include designed experiments in agriculture and microbiology, where entire patterns or images of plants on plots of land or bacteria in Petri dishes (rather than only their total counts) could be analyzed under combinations of treatments.

APPENDIX A: AUXILIARY RESULTS USED FOR THE PROOF OF THEOREM 2

For completeness and self-containedness, we state here (consequences of) results from the literature as well as some additional calculations needed for the proof of Theorem 2.

First, we formulate a straightforward generalization of Hoeffding's theorem for the asymptotic normality of U-statistics (univariate version of Theorem 7.1 in [28]) for random elements in the general metric space $\mathcal{X}$ with countably generated Borel σ-algebra. See also Theorem 1(b) of [15], where this result is further generalized to (weakly) dependent sequences of random elements.

THEOREM 6. *Let $(X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence of $\mathcal{X}$-valued random elements. Let $h\colon \mathcal{X}^m \to \mathbb{R}$ be symmetric and nondegenerate in the sense that there are $x_2, \dots, x_m \in \mathcal{X}$ such that*

$$\mathbb{E}h(X_1, x_2, \dots, x_m) \neq 0.$$

Suppose further that $\mathbb{E}(h(X_1, \dots, X_m)^2) < \infty$. We write

$$U_n = \binom{n}{m}^{-1} \sum_{\substack{i_1,\dots,i_m=1\\ i_1<\cdots<i_m}}^{n} h(X_{i_1}, \dots, X_{i_m}).$$

for the U-statistic with kernel h. Then

$$\sqrt{n}\big(U_n - \mathbb{E}(U_n)\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, m^2\gamma_h^2),$$

where for an independent copy $(\tilde{X}_2, \dots, \tilde{X}_m)$ of $(X_2, \dots, X_m)$

$$\begin{aligned}\gamma_h^2 &= \mathrm{Cov}\big(h(X_1, X_2, \dots, X_m), h(X_1, \tilde{X}_2, \dots, \tilde{X}_m)\big)\\ &= \mathrm{Var}\big(\mathbb{E}\big(h(X_1, \dots, X_m)|X_1\big)\big).\end{aligned}$$

REMARK 7. In the setting of Theorem 6 above, Theorem 5.2 of [28] yields

$$m^2\gamma_h^2 \le n\,\mathrm{Var}(U_n) \le m\,\mathrm{Var}\big(h(X_1, \dots, X_m)\big)$$

for all $n \ge m$. The right-hand bound is sharp for $n = m$ and $n\,\mathrm{Var}(U_n) \searrow m^2\gamma_h^2$ as $n \to \infty$.

The above inequality means in particular that for finite n the expression $\frac{m^2}{n}\gamma_h^2$ can only underestimate $\mathrm{Var}(U_n)$. The exact formula for $m = 2$ is

$$n\,\mathrm{Var}(U_n) = \frac{n-2}{n-1} \cdot 4\gamma_h^2 + \frac{1}{n-1} \cdot 2\,\mathrm{Var}\big(h(X_1, X_2)\big).$$

The next result is similar to classical ANOVA. For completeness, we give its proof.

LEMMA 8. *Let $C \in \mathbb{R}^{(k-1)\times k}$ as in (12), $D \in \mathbb{R}^{n\times k}$ as in (11), $U = (u_1, \dots, u_k)'$ and $\nu = (n_1, \dots, n_k)'$. We have*

$$C'\big(C(D'D)^{-1}C'\big)^{-1}C = D'D - \frac{1}{n}\nu\nu'$$

and

$$U'\Big(D'D - \frac{1}{n}\nu\nu'\Big)U = \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (u_i - u_j)^2$$

PROOF. Define

$$\nu_{(i)} := (n_1, \dots, n_i)', \quad \Lambda_{(i)} := \mathrm{diag}(\nu_{(i)}) \in \mathbb{R}^{i\times i} \quad \text{and}$$

$$\mathbb{1}_{(i)} := (1, \dots, 1)' \in \mathbb{R}^{i}.$$

Then $\mathbb{1}_{(i)}\mathbb{1}_{(i)}'$ is the $i \times i$ matrix of 1's. We build up the equality step by step. Since $D'D = \Lambda_{(k)}$ and, therefore,

$$\big(D'D\big)^{-1} = (\Lambda_{(k)})^{-1} = \mathrm{diag}(1/n_1, \dots, 1/n_k),$$

We obtain

$$C\big(D'D\big)^{-1}C' = (\Lambda_{(k-1)})^{-1} + \frac{1}{n_k} \cdot \mathbb{1}_{(k-1)}\mathbb{1}_{(k-1)}'$$

and

$$\big(C\big(D'D\big)^{-1}C'\big)^{-1} = \Lambda_{(k-1)} - \frac{1}{n} \cdot \nu_{(k-1)}\nu_{(k-1),}'$$

and finally

$$C'\big(C\big(D'D\big)^{-1}C'\big)^{-1}C = \Lambda_{(k)} - \frac{1}{n} \cdot \nu\nu'$$

When we multiply the vector U from left and right, the ijth entry in the matrix is the coefficient of u_iu_j. This leads to

$$\begin{aligned}
&U'\Big(D'D - \frac{1}{n}\nu\nu'\Big)U\\
&= \sum_{i=1}^{k} n_i u_i^2 - \frac{1}{n}\sum_{i=1}^{k} n_i^2 u_i^2 - \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} 2n_i n_j u_i u_j\\
&= \frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 \sum_{j=1}^{k} n_j - \frac{1}{2n}\sum_{i=1}^{k} n_i^2\big(u_i^2 + u_i^2\big)\\
&\quad - \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} 2n_i n_j u_i u_j\\
&= \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j\big(u_i^2 + u_j^2\big) - \frac{1}{2n}\sum_{i=1}^{k} n_i^2\big(u_i^2 + u_i^2\big)\\
&\quad - \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} 2n_i n_j u_i u_j\\
&= \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j\big(u_i^2 + u_j^2\big) - \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} 2n_i n_j u_i u_j\\
&= \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (u_i - u_j)^2.
\end{aligned}$$

□

REMARK 9. Let $\bar{u} = \frac{1}{k}\sum_{i=1}^{k} u_i$. An equivalent expression for $U'(D'D - \frac{1}{n}\nu\nu')U$ in Lemma 8 can be computed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (u_i - u_j)^2 \\
&\quad= \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j \big((u_i - \bar{u}) + (\bar{u} - u_j)\big)^2 \\
&\quad= \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j (u_i - \bar{u})^2 \\
&\qquad+ \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j (\bar{u} - u_j)^2 \\
&\qquad+ \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j (u_i - \bar{u})(\bar{u} - u_j) \\
&\quad= \sum_{i=1}^{k} n_i (u_i - \bar{u})^2 + \frac{1}{2n}\sum_{i=1}^{k} n_i (u_i - \bar{u}) \sum_{j=1}^{k} n_j (\bar{u} - u_j) \\
&\quad= \sum_{i=1}^{k} n_i (u_i - \bar{u})^2 + \frac{1}{2n}\left(\sum_{i=1}^{k} n_i (u_i - \bar{u})\right)^2
\end{aligned}$$

If $n_1 = \cdots = n_k = \tilde{n}$, we see directly from the right-hand side that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (u_i - u_j)^2 &= \sum_{i=1}^{k} n_i (u_i - \bar{u})^2 \\
&= \tilde{n}\sum_{i=1}^{k} (u_i - \bar{u})^2.
\end{aligned}$$

The following is a well-known lemma about the distribution of quadratic forms, proved in [22], Theorem 2. Denote by $\chi_r^2(\lambda)$ the noncentral χ^2-distribution with r degrees of freedom and noncentrality parameter λ. By definition, this is the distribution of $\sum_{i=1}^{r} Y_i^2$ for any independent $Y_i \sim \mathcal{N}(\mu_i, 1)$ with $\lambda = \sum_{i=1}^{r} \mu_i^2$ (note that we use a slightly different parametrization than [22]).

LEMMA 10. *Let $Z \sim \mathcal{N}_k(\mu, I)$ for some $\mu \in \mathbb{R}^k$ and let $G \in \mathbb{R}^{k\times k}$ be symmetric and idempotent. Then $Z'GZ \sim \chi_r^2(\lambda)$, where $r = \operatorname{trace}(G) = \operatorname{rank}(G)$ and $\lambda = \mu'\mu$.*

APPENDIX B: SIMULATION STUDY FOR THE CONVERGENCE SPEED IN THEOREM 2

In the present subsection, we numerically assess the speed of convergence of our new $\widetilde{L}$ statistic under the null hypothesis of equal group distributions toward the χ_{k-1}^2 distribution as presented in Section 4.2. For comparison, we also consider the Fréchet T_L and T statistics, which were shown in [16] to have a limiting χ_{k-1}^2 distribution as well.

Our experiments are based on $k = 2$ groups, both simulated from the same distribution, which is either CSR(35) or the Strauss hard core distribution with $\lambda = 35$. These are the extreme distributions having either no interaction or very strong interaction in Section 6.2. As group size, we consider $\tilde{n} = 5, 20, 50, 200$. The computation of the Fréchet T and T_L depend on the calculation of a barycenter. For this, we used the heuristic algorithm presented in [35]. The calculation of an exact barycenter is computationally infeasible for this kind of data. To compensate that we do not get the optimal solution, we did 5 restarts in every barycenter calculation and used the best of the 5 solutions as the barycenter.

Block A of Figure 9 shows QQ-plots for the empirical distributions of our new Levene statistic $\widetilde{L}$, the Fréchet statistic T_L and the Fréchet statistic T on the y-axis and the theoretical χ_1^2 distribution on the x-axis. The data are the CSR(35) point patterns. In the first column, the groups consist of $\tilde{n} = 5$ patterns, in the second column of $\tilde{n} = 20$ patterns and so on. For the two Levene statistics $\widetilde{L}$ and T_L, we can see the computed quantiles approach the theoretical quantiles as the group size $\tilde{n}$ gets larger. Even for a medium group size $\tilde{n} = 50$ the computed quantiles are very close to the theoretical quantiles of a χ_1^2 distribution.

Similarly, Block B of Figure 9 shows QQ-plots for hardcore Strauss distributed point patterns. Again the four columns correspond to the four group sizes $\tilde{n} = 5, 20, 50, 200$ and the three rows correspond to the three statistics. For this data, the computed quantiles are already very close to the theoretical quantiles of a χ_1^2 distribution for $\tilde{n} = 20$ for the two Levene statistics.

In both cases, the third row, the combined Fréchet statistic T, yields quantiles that are far from the theoretical quantiles. This is solely due to the summand T_F that is not considered in the second row.

ACKNOWLEDGMENTS

The authors would like to thank an anonymous referee for thoughtful comments that led to an improvement of contents and presentation.

FUNDING

Raoul Müller was supported by Deutsche Forschungsgemeinschaft GRK 2088. Jorge Mateu was partially supported by projects PID2019-107392RB-I00 from the Spanish Ministry of Science and by UJI-B2021-37 from University Jaume I.

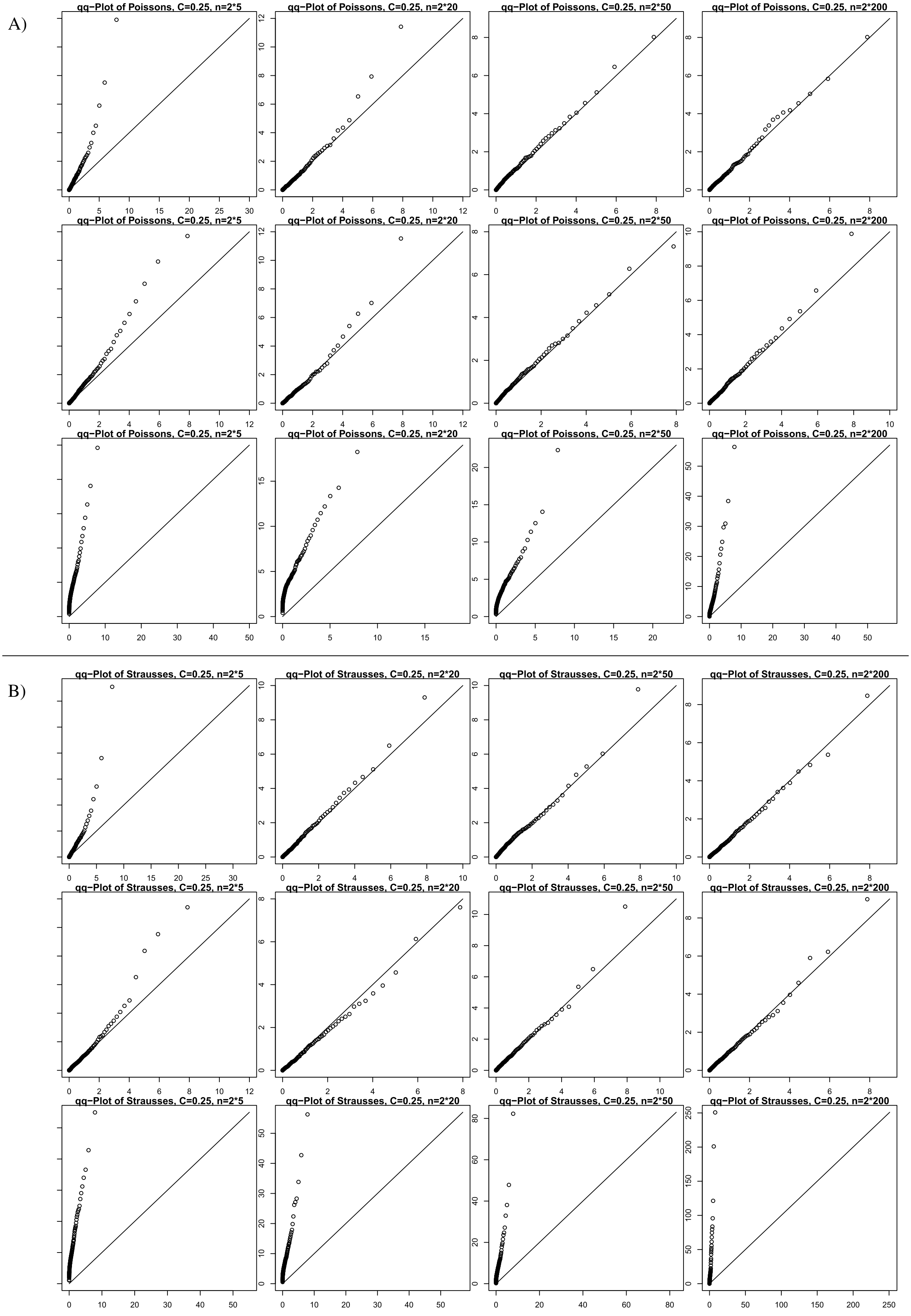

FIG. 9. *QQ-plots of the percentiles based on* 500 *statistics values* (*on the* y-*axis*) *versus* χ_1^2-*percentiles*. *Based on* $k = 2$ *groups of* $\tilde{n} = 5, 20, 50, 200$ *patterns from* CSR(35) (*Block A*, *top*) *and from a Strauss hard core distribution with* $\lambda = 35$ (*Block B*, *bottom*). *In either block*, *the first row is our new* $\widetilde{L}$ *statistic* (7), *the second and third rows are the Fréchet* T_L *statistic from Section* 3.2 *and the Fréchet* T *statistic*, *respectively*.

REFERENCES

[1] ALEKSEYENKO, A. V. (2016). Multivariate Welch t-test on distances. *Bioinformatics* **32** 3552–3558. https://doi.org/10.1093/bioinformatics/btw524

[2] ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

[3] ANDERSON, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62** 245–253. MR2226579 https://doi.org/10.1111/j.1541-0420.2005.00440.x

[4] ANDERSON, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). *Wiley Statsref*: *Statistics Reference Online* 1–15.

[5] ANDERSON, M. J., WALSH, D. C. I., CLARKE, K. R., GORLEY, R. N. and GUERRA-CASTRO, E. (2017). Some solutions to the multivariate Behrens–Fisher problem for dissimilarity-based analyses. *Aust. N. Z. J. Stat.* **59** 57–79. MR3635167 https://doi.org/10.1111/anzs.12176

[6] BERTSEKAS, D. P. (1988). The auction algorithm: A distributed relaxation method for the assignment problem. *Ann. Oper. Res.* **14** 105–123. MR0963896 https://doi.org/10.1007/BF02186476

[7] BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

[8] BORGWARDT, S. and PATTERSON, S. (2020). Improved linear programs for discrete barycenters. *INFORMS J. Optim.* **2** 14–33. MR4172566 https://doi.org/10.1287/ijoo.2019.0020

[9] BORGWARDT, S. and PATTERSON, S. (2021). On the computational complexity of finding a sparse Wasserstein barycenter. *J. Comb. Optim.* **41** 736–761. MR4228512 https://doi.org/10.1007/s10878-021-00713-5

[10] BROWN, M. B. and FORSYTHE, A. B. (1974). Robust tests for the equality of variances. *J. Amer. Statist. Assoc.* **69** 364–367.

[11] BROWN, M. B. and FORSYTHE, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics* **16** 129–132. MR0334368 https://doi.org/10.2307/1267501

[12] CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2004). An anova test for functional data. *Comput. Statist. Data Anal.* **47** 111–122. MR2087932 https://doi.org/10.1016/j.csda.2003.10.021

[13] DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I*: *Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

[14] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II*: *General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

[15] DENKER, M. and KELLER, G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Z. Wahrsch. Verw. Gebiete* **64** 505–522. MR0717756 https://doi.org/10.1007/BF00534953

[16] DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821. MR4031200 https://doi.org/10.1093/biomet/asz052

[17] FISHER, R. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

[18] GASTWIRTH, J. L., GEL, Y. R. and MIAO, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statist. Sci.* **24** 343–360. MR2757435 https://doi.org/10.1214/09-STS301

[19] GE, D., WANG, H., XIONG, Z. and YE, Y. (2019). Interior-point methods strike back: Solving the Wasserstein barycenter problem. *Adv. Neural Inf. Process. Syst.* **32**.

[20] GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. MR3693544 https://doi.org/10.1214/16-AOAS1015

[21] GONZÁLEZ, J. A., LAGOS-ÁLVAREZ, B. M. and MATEU, J. (2021). Two-way layout factorial experiments of spatial point pattern responses in mineral flotation. *TEST* **30** 1046–1075. MR4346817 https://doi.org/10.1007/s11749-021-00768-w

[22] GRAYBILL, F. A. and MARSAGLIA, G. (1957). Idempotent matrices and quadratic forms in the general linear hypothesis. *Ann. Math. Stat.* **28** 678–686. MR0092307 https://doi.org/10.1214/aoms/1177706879

[23] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[24] HAMIDI, B., WALLACE, K., VASU, C. and ALEKSEYENKO, A. V. (2019). W_d^*-Test: Robust distance-based multivariate analysis of variance. *Microbiome* **7** 1–9.

[25] HEINEMANN, F. (2021). WSGeometry: Geometric Tools Based on Balanced/Unbalanced Optimal Transport. R package version 1.2.1. Available at https://CRAN.R-project.org/package=WSGeometry.

[26] HEINEMANN, F., KLATT, M. and MUNK, A. (2023). Kantorovich–Rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *Appl. Math. Optim.* **87** Paper No. 4. MR4506758 https://doi.org/10.1007/s00245-022-09911-x

[27] HEINEMANN, F., MUNK, A. and ZEMEL, Y. (2022). Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM J. Math. Data Sci.* **4** 229–259. MR4386483 https://doi.org/10.1137/20M1385263

[28] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294 https://doi.org/10.1214/aoms/1177730196

[29] HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. Technical Report, Mimeograph Series No. 302. Dept. Statistics, Univ. North Carolina.

[30] HUCKEMANN, S., HOTZ, T. and MUNK, A. (2009). Intrinsic MANOVA for Riemannian manifolds with an application to Kendall's space of planar shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 593–603.

[31] LEE, Y. T. and SIDFORD, A. (2014). Path-finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014* 424–433. IEEE Computer Soc., Los Alamitos, CA. MR3344892 https://doi.org/10.1109/FOCS.2014.52

[32] LEVENE, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics. Stanford Studies in Mathematics and Statistics* **2** 278–292. Stanford Univ. Press, Stanford, CA. MR0120709

[33] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. *Probability and Mathematical Statistics*: *A Series of Monographs and Textbooks*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto. MR0560319

[34] MÜLLER, R. and SCHUHMACHER, D. (2019–2022). ttbary: Barycenter Methods for Spatial Point Patterns. R package version 0.3-0. Available at https://CRAN.R-project.org/package=ttbary.

[35] MÜLLER, R., SCHUHMACHER, D. and MATEU, J. (2020). Metrics and barycenters for point pattern data. *Stat. Comput.* **30** 953–972. MR4108686 https://doi.org/10.1007/s11222-020-09932-y

[36] RAMÓN, P., DE LA CRUZ, M., CHACÓN-LABELLA, J. and ESCUDERO, A. (2016). A new non-parametric method for analyzing replicated point patterns in ecology. *Ecography* **39** 1109–1117.

[37] RIZZO, M. L. and SZÉKELY, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 https://doi.org/10.1214/09-AOAS245

[38] SCHEFFÉ, H. (1967). *The Analysis of Variance*, 1st ed. John Wiley & Sons.

[39] SCHUHMACHER, D., BÄHRE, B., BONNEEL, N., GOTTSCHLICH, C., HARTMANN, V., HEINEMANN, F., SCHMITZER, B. and SCHRIEBER, J. (2014–2022). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.13-0. Available at https://CRAN.R-project.org/package=transport.

[40] SONG, H. and CHEN, H. (2022). New graph-based multi-sample tests for high-dimensional and non-Euclidean data. Preprint. Available at https://arxiv.org/abs/2205.13787.

[41] TAMAYO-URIA, I., MATEU, J. and DIGGLE, P. J. (2014). Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spat. Stat.* **9** 192–206. MR3326839 https://doi.org/10.1016/j.spasta.2014.03.005

[42] WELCH, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika* **38** 330–336. MR0046617 https://doi.org/10.1093/biomet/38.3-4.330

[43] WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

[44] ZHANG, J.-T., GUO, J. and ZHOU, B. (2022). Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *J. Econometrics*. https://doi.org/10.1016/j.jeconom.2022.03.007.

[45] ZHANG, Q., MAHDI, G., TINKER, J. and CHEN, H. (2020). A graph-based multi-sample test for identifying pathways associated with cancer progression. *Comput. Biol. Chem.* **87** 107285.