Received: 4 April 2020 | Accepted: 4 February 2022

DOI: 10.1111/rssc.12553

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series C Applied Statistics

Assessing seismic origin of geological features by fitting equidistant parallel lines

P.E. Jupp[1] | **I.B.J. Goudie**[1] | **R.A. Batchelor**[1,†] | **R.J.B. Goudie**[2]

[1]University of St Andrews, St Andrews, UK

[2]University of Cambridge, Cambridge, UK

Correspondence
P.E. Jupp, School of Mathematics and Statistics, North Haugh, St Andrews, Fife KY16 9SS, UK.
Email: pej@st-andrews.ac.uk

Abstract

Some planes in sedimentary rocks contain features that appear to lie near equally spaced parallel lines. Determining whether or not they do so can provide information on possible mechanisms for their formation. The problem is recast here in terms of circular statistics, enabling closeness of candidate sets of lines to the points to be measured by a mean resultant length. This leads to a test of goodness of fit and to estimates of the direction of the lines and of the spacing between them. Two contrasting data sets are analysed.

KEYWORDS

directional statistics, quantal model

1 | INTRODUCTION

Geological features that are points or line segments and lie near almost-parallel straight lines occur in various contexts in the earth sciences. Examples include corrugations on fault surfaces (Resor & Meer, 2009), fault lines in the Earth's crust and magma dyke swarms. One class of such features is that in which (a) each feature consists of a set of points near some straight line, (b) interest lies in whether or not these straight lines are equally spaced. Some intriguing groups of features of this kind are found in bedding planes in Lower Carboniferous sediments. It was these that motivated the work described here. If it can be shown that there is support for the assertion that these features tend to lie near equally spaced parallel straight lines then this can be regarded as evidence that they were formed as the result of seismic activity.

[†]R.A. Batchelor died on 15th February 2022.

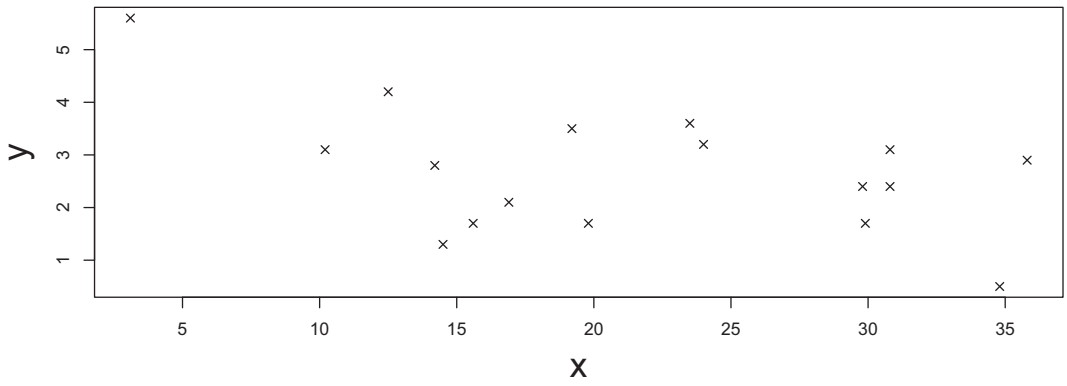

FIGURE 1 Positions of raised sedimentary features in Lower Carboniferous sediment at Cellardyke, Fife, UK. Units of measurement of axes in metres

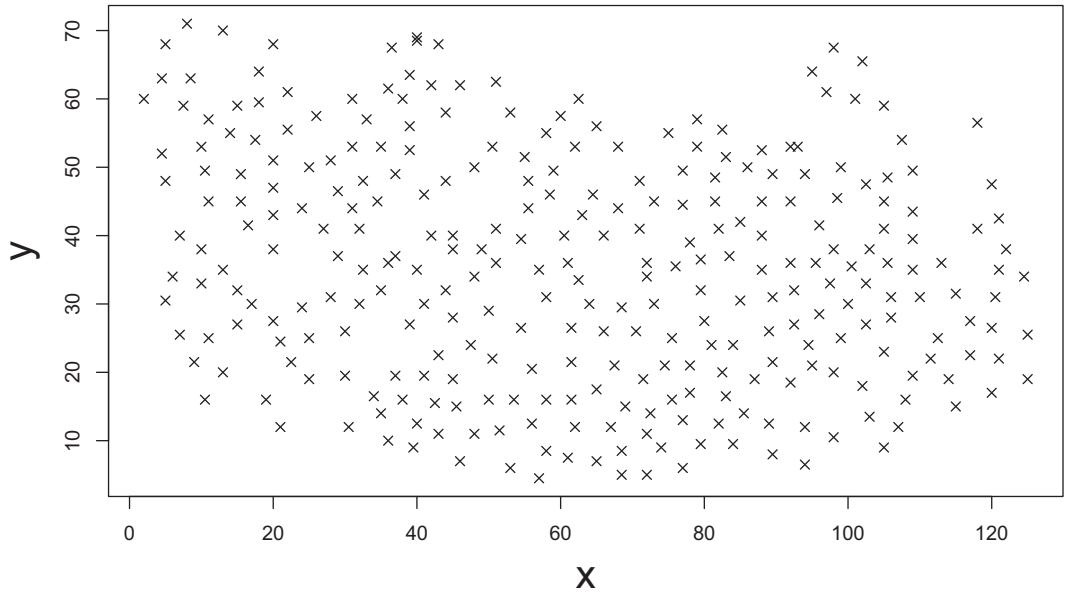

FIGURE 2 Positions of saucer-shaped depressions in Lower Carboniferous sediment at Catcraig, East Lothian, UK. Data from Batchelor et al. (submitted). Units of measurement of axes in metres

Two data sets of interest here are (a) 17 locations of raised sedimentary features in a bedding plane at Cellardyke, Fife, UK, shown in Figure 1, and (b) 300 locations of saucer-shaped depressions in a bedding plane at Catcraig, East Lothian, UK, shown in Figure 2. Data set (b) is from Batchelor et al. (submitted). The way in which the depressions were formed has long been controversial. They were popularly believed to be organic in origin, which would be expected to imply that their locations are uniformly distributed. The analysis of this data set in Section 3.2 indicates that, rather than this being the case, they are aligned near parallel lines. The alignment indicates to the authors of Batchelor et al. (submitted) that an underlying geophysical process was involved, which is interpreted as the result of sediment liquefaction caused by seismic waves generated during an earthquake.

In the present paper we give a general formalisation of near linearity of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a plane and we introduce a method of investigating the conjecture that

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \text{ lie 'near' equally spaced parallel lines in the plane.} \tag{1}$$

Conjecture (1) is reminiscent of the conjecture about given (real number) observations $y_1, \ldots, y_n$ that

$$y_1, \ldots, y_n \text{ are 'nearly' multiples of } q \tag{2}$$

for some unknown fixed *quantum* q. A slightly weaker conjecture is

$$\text{the differences between } y_1, \ldots, y_n \text{ are 'nearly' multiples of } q. \tag{3}$$

Conjectures (2) and (3) are assessed using *quantal models*. An early instance of conjecture (2) in which q is known (considered before the discovery of isotopes) concerned whether or not atomic weights are almost integers (von Mises, 1918); see Mardia and Jupp (2000), Example 6.4. Another important instance, considered in detail in Kendall (1974), arose from the conjecture that key distances in megalithic stone structures in the British Isles are almost multiples of a hypothesised unit of length, the *megalithic yard*. A Bayesian analysis of this problem was given by Freeman (1976). Pakkanen (2002) applied Kendall's methods to the detection and estimation of a standard length ('foot') in ancient Greek architecture. Further examples in archaeology, biology and cosmology are discussed in the historical survey part of Çankaya and Fieller (2009).

In Section 2 we propose a semi-parametric statistical model for points in a plane which appear to lie near equally spaced parallel lines. It is similar in spirit to one of the standard quantal models used to handle conjecture (2). A parametric sub-model is explored. Techniques from directional statistics can be used (a) to test whether or not the points do indeed lie close to a suitable set of parallel lines (as in conjecture (1)), (b) to estimate the spacing and direction of the best-fitting set of parallel lines, (c) to assess the goodness of fit of the lines (estimated using the parametric sub-model) to the observed points. The model assumes spatial homogeneity across the region of the plane in which the points lie. A method of assessing whether or not the data support this assumption is given in Section 2.6. Section 2.7 considers inference for the case in which the data are measured only up to limited resolution. A small simulation study is given in Section 2.8. Section 3 illustrates the methods by analysing the Cellardyke and Catcraig data sets of Figures 1 and 2 respectively. Section 4 gives some concluding remarks.

2 | A STATISTICAL MODEL

Any set of parallel lines in the plane determines an axis $\pm\mathbf{u}$ normal to the lines. Here $\mathbf{u}$ is a direction (unit vector). The ambiguity of sign can be removed by transforming $\pm\mathbf{u}$ (where $\mathbf{u}$ is treated as a column vector) to its 'square', the symmetric 2×2 matrix $\mathbf{u}\mathbf{u}^\top$. This is the standard way of handling axes and is used, for example, in the Bingham distributions (see Section 9.4.3 of Mardia & Jupp, 2000). For computational purposes it is convenient to restrict $\mathbf{u}$ to lie in some given semi-circle.

2.1 | The general model

A reasonable statistical model in our context is that $\mathbf{u}^\top\mathbf{x}_1, \ldots, \mathbf{u}^\top\mathbf{x}_n$ are independent observations on real random variables $Y_1, \ldots, Y_n$ that satisfy

$$Y_i = \beta + m_i q + e_i \quad i = 1, \ldots n, \tag{4}$$

for some quantum q, where $m_1, \ldots, m_n$ are unknown integers. Conjecture (1) can be formalised as the hypothesis

$$\mathbf{u}^\top\mathbf{x}_1, \ldots, \mathbf{u}^\top\mathbf{x}_n \text{ are generated by model (4).} \tag{5}$$

Although, in principle, the quantum q in conjectures (2)–(3) and Equation (4) can take any positive value, when estimating q it is necessary to restrict q to values that are neither too small (e.g. so that each point lies near its 'own' line) nor too large (e.g. so that all the points are contained within a single pair of parallel lines). Allowing such extreme values of q would lead to an estimate of unrealistic size. We shall take q in $(q_{\min}, q_{\max})$ for some suitable positive $q_{\min}$ and $q_{\max}$. Without loss of generality we can take the parameter β in Equation (4) to satisfy $-q/2 \le \beta < q/2$ and assume that the random variables $e_1, \ldots, e_n$ tend to be near 0. Model (4) is the 'shifted quantal model' (2) of Çankaya and Fieller (2009).

It is useful to combine $\mathbf{u}$ and q into the vector $\mathbf{v}$, where

$$\mathbf{v} = 2\pi q^{-1}\mathbf{u}. \tag{6}$$

For $i = 1, \ldots, n$, put

$$\theta_i(\mathbf{v}) = 2\pi \left\{\mathbf{v}^\top\mathbf{x}_i/(2\pi)\right\}, \tag{7}$$

where $\{\cdot\}$ denotes the fractional part, as it does henceforth. Then $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ lie in $[0, 2\pi)$, and so can be considered as angles representing points on the circle of unit radius. If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ lie near parallel lines normal to $\mathbf{u}$ and distance q apart then $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ will be near $2\pi\beta/q$ on the circle. On the other hand, if the components of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ along $\mathbf{u}$ are more or less uniformly spread over some interval of length much greater than q then $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ will resemble observations from the uniform distribution on the circle.

The concentration of the angles $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ is conveniently measured by the *mean resultant length*, $\overline{R}(\mathbf{v})$, defined as

$$\overline{R}(\mathbf{v}) = \left[\overline{C}(\mathbf{v})^2 + \overline{S}(\mathbf{v})^2\right]^{1/2}, \tag{8}$$

where

$$\overline{C}(\mathbf{v}) = n^{-1}\sum_{i=1}^{n}\cos\theta_i(\mathbf{v}), \qquad \overline{S}(\mathbf{v}) = n^{-1}\sum_{i=1}^{n}\sin\theta_i(\mathbf{v}). \tag{9}$$

A value of $\overline{R}(\mathbf{v})$ near 1 indicates that the angles are concentrated, whereas a value near 0 occurs when the angles are almost uniformly spread around the circle (or, more generally, display some antipodal symmetry). The location of $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ can be described by the *mean direction* $\hat{\mu}_{\mathbf{v}}$, given by

$$\overline{C}(\mathbf{v}) = \overline{R}(\mathbf{v})\cos\hat{\mu}_{\mathbf{v}}, \qquad \overline{S}(\mathbf{v}) = \overline{R}(\mathbf{v})\sin\hat{\mu}_{\mathbf{v}}. \tag{10}$$

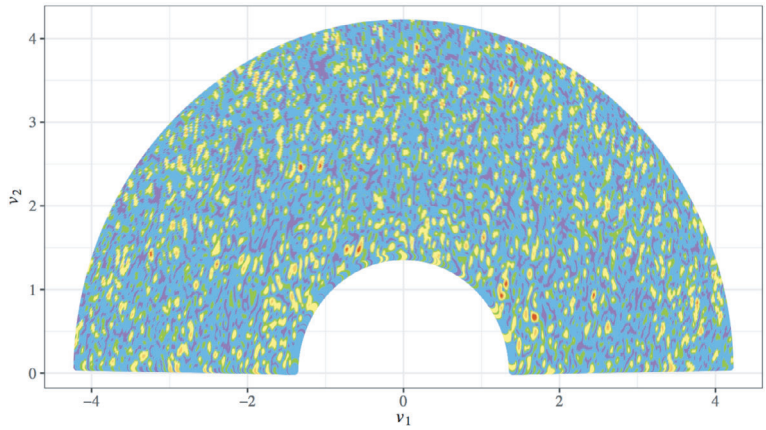

FIGURE 3 Heat map of the mean resultant length $\overline{R}(\mathbf{v})$ for the Catcraig data set as a function of $\mathbf{v} = (v_1, v_2)$, with the boundaries between red, orange, yellow, green, blue and purple at multiples of 0.75, 0.6, 0.4, 0.3 and 0.1, respectively, of the maximum of $\overline{R}(\mathbf{v})$

2.2 | Estimation

It is intuitively reasonable to estimate $\mathbf{v}$ by $\hat{\mathbf{v}}$, which brings the equally spaced parallel lines as close as possible to the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$; more precisely,

$$\hat{\mathbf{v}} \text{ maximises } \overline{R}(\mathbf{v}) \text{ over all vectors } \mathbf{v}, \text{ with } 2\pi\|\mathbf{v}\|^{-1} \in (q_{\min}, q_{\max}). \tag{11}$$

The parameter β in Equation (4) can then be estimated by

$$\hat{\beta} = \|\hat{\mathbf{v}}\|^{-1}\, \hat{\mu}_{\hat{\mathbf{v}}}. \tag{12}$$

A little algebra (like that used to obtain (3.8) of Pewsey et al., 2013) shows that

$$\overline{R}(\mathbf{v}) = n^{-1} \sum_{i=1}^{n} \cos[\theta_i(\mathbf{v}) - \hat{\mu}_{\mathbf{v}}]. \tag{13}$$

It follows from (7), (9), (10) and (13) that $\overline{R}(\mathbf{v})$ is a smooth (i.e. infinitely differentiable) function of $\mathbf{v}$. In spite of this smoothness, the function can appear very spiky, as is evident from both the heat map in Figure 3 and the three-dimensional plot in Figure 4 of the mean resultant length $\overline{R}(\mathbf{v})$ for the Catcraig data set. The spikiness means that iterative methods of obtaining $\hat{\mathbf{v}}$ that are based on derivatives will succeed only if initiated from a large number of initial values located at the points of a fine grid.

It is useful to define the i-th *residual* as

$$r_i = \theta_i(\hat{\mathbf{v}}) - \hat{\mu}_{\hat{\mathbf{v}}} \pmod{2\pi} \quad i = 1, \ldots, n, \tag{14}$$

where the reduction mod 2π is chosen to ensure that r_i lies in $[-\pi, \pi)$.

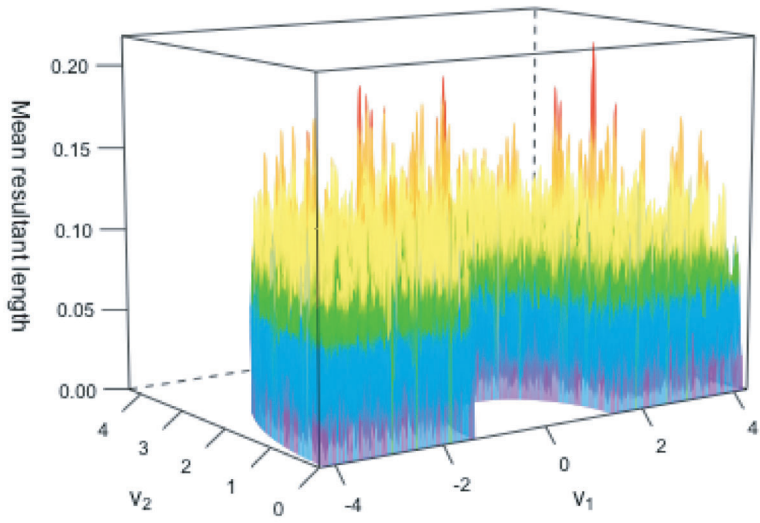

FIGURE 4 Three-dimensional plot of the mean resultant length $\overline{R}(\mathbf{v})$ for the Catcraig data set as a function of $\mathbf{v} = (v_1, v_2)$, with the colours and boundaries matching those in Figure 3

The i-th *fitted point*, $\hat{\mathbf{x}}_i$, is the projection of the observed point $\mathbf{x}_i$ along $\hat{\mathbf{u}}$ onto the nearest fitted line. Thus

$$\hat{\mathbf{x}}_i = \mathbf{x}_i - (\hat{q}/2\pi) r_i \hat{\mathbf{u}}. \tag{15}$$

2.3 | The von Mises case

Von Mises (1918) showed that, under mild conditions, the maximum likelihood estimate of the location parameter μ in a distribution on the unit circle is the sample mean direction if and only if the distribution is the one that he introduced and that now bears his name. A circular random variable has the von Mises distribution $M(\mu, \kappa)$ with mean direction μ and concentration κ if it has density

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp[\kappa \cos(\theta - \mu)], \quad 0 \le \theta \le 2\pi, \tag{16}$$

I_ν denoting the modified Bessel function of the first kind and order ν. Details of von Mises distributions and of the above characterisation of them are discussed in Mardia and Jupp (2000), Section 3.5.4. See also Fisher (1993), Section 3.3.6 and Jammalamadaka and SenGupta (2001), Section 2.2.4.

It follows from von Mises's characterisation that, under rather weak conditions, the class of parametric models for $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ for which the maximum likelihood estimates of v are the moment estimates given by definition (11) consists only of von Mises distributions. More precisely, if (a) $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ are independent observations from a density on the circle of the form $f(\theta; \mu) = g(\theta - \mu)$ for some positive function g with continuous second derivative, (b) for $n = 2, 3$, the maximum likelihood estimate of μ is the sample mean direction, as defined in Equations (10), then

$$\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v}) \text{ are independent observations from } M(\mu, \kappa) \tag{17}$$

for some positive κ. Note that condition (b) needs be satisfied only for $n = 2$ and $n = 3$ for the resulting von Mises property to hold, as is shown in Mardia and Jupp (2000) Section 3.5.4.

In view of Equations (4), (6), (7) and hypothesis (5), it is appropriate to take

$$\mu = 2\pi\,\beta/q. \tag{18}$$

Model (17) is analogous to the model used in Çankaya and Fieller (2009) to assess whether or not scalar observations $y_1, \ldots, y_n$ satisfy (2) or (3). On the region $\kappa > 0$, the parameters $\pm(\mathbf{v}, \mu)$, κ in the model given by Equation (7) and assumption (17) are identifiable, and so are estimated consistently by their maximum likelihood estimators. If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent then the joint density of $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ is

$$f(\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v}); \mathbf{v}, \mu, \kappa) = \frac{1}{[2\pi I_0(\kappa)]^n} \exp\left[\kappa \sum_{i=1}^{n} \cos\left(\theta_i(\mathbf{v}) - \mu\right)\right], \tag{19}$$

where $0 \le \theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v}) \le 2\pi$.

Assumption (17) of an underlying von Mises distribution can be tested by applying the test of 'von Misesness' based on Watson's U^2 test of uniformity (Mardia & Jupp, 2000, Sections 6.3.3, 6.4.2) to $\theta_1(\hat{\mathbf{v}}), \ldots, \theta_n(\hat{\mathbf{v}})$. Small values of U^2 indicate a good fit.

Under assumption (17), it follows from (13) and (19) that the log-likelihood of the parameters $\mathbf{v}$, μ, κ based on the angles $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ is

$$\ell(\mathbf{v}, \mu, \kappa; \theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})) = n\kappa\overline{R}(\mathbf{v})\cos(\hat{\mu}_{\mathbf{v}} - \mu) - n\log I_0(\kappa). \tag{20}$$

For fixed $\mathbf{v}$, the maximum of expression (20) is at $(\mu, \kappa) = (\hat{\mu}_{\mathbf{v}}, \hat{\kappa}_{\mathbf{v}})$ with

$$A(\hat{\kappa}_{\mathbf{v}}) = \overline{R}(\mathbf{v}), \tag{21}$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$. Thus the profile log-likelihood of $\mathbf{v}$ is

$$\begin{aligned} \ell_P(\mathbf{v}) &= \ell(\mathbf{v}, \hat{\mu}_{\mathbf{v}}, \hat{\kappa}_{\mathbf{v}}; \theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})) \\ &= n\hat{\kappa}_{\mathbf{v}}\overline{R}(\mathbf{v}) - n\log I_0(\hat{\kappa}_{\mathbf{v}}) \\ &= n\hat{\kappa}_{\mathbf{v}}A(\hat{\kappa}_{\mathbf{v}}) - n\log I_0(\hat{\kappa}_{\mathbf{v}}). \end{aligned} \tag{22}$$

Differentiation of Equation (22) with respect to $\hat{\kappa}_{\mathbf{v}}$ gives

$$\frac{d\ell_P(\mathbf{v})}{d\hat{\kappa}_{\mathbf{v}}} = n\{A(\hat{\kappa}_{\mathbf{v}}) + \hat{\kappa}_{\mathbf{v}}A'(\hat{\kappa}_{\mathbf{v}}) - A(\hat{\kappa}_{\mathbf{v}})\} = n\hat{\kappa}_{\mathbf{v}}A'(\hat{\kappa}_{\mathbf{v}}),$$

and so, using the derivative of Equation (21),

$$\frac{d\ell_P(\mathbf{v})}{d\overline{R}(\mathbf{v})} = \frac{d\ell_P(\mathbf{v})}{d\hat{\kappa}_{\mathbf{v}}} \Big/ \frac{d\overline{R}(\mathbf{v})}{d\hat{\kappa}_{\mathbf{v}}} = n\hat{\kappa}_{\mathbf{v}} \ge 0.$$

Thus

$$\ell_P(\mathbf{v}) \text{ is an increasing function of } \overline{R}(\mathbf{v}). \tag{23}$$

It follows that the maximum likelihood estimate of $\mathbf{v}$ in the von Mises model is the same as the estimate $\hat{\mathbf{v}}$ given in definition (11). The maximum likelihood estimates $\hat{\beta}$ and $\hat{\kappa}$ of β and κ are given by Equation (12) and

$$A(\hat{\kappa}) = \hat{R}, \tag{24}$$

where

$$\hat{R} = \overline{R}(\hat{\mathbf{v}}). \tag{25}$$

The scenario of $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ resembling observations from the uniform distribution can be formalised under assumption (17) as the null hypothesis

$$H_0 : \kappa = 0,$$

whereas the scenario of $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ being clustered near the point μ on the circle can be formalised as the alternative hypothesis

$$H_1 : \kappa > 0.$$

The likelihood ratio test of H_1 versus H_0 rejects H_0 for large values of the maximised log-likelihood, $\ell_P(\hat{\mathbf{v}})$. Since $\ell_P(\mathbf{v})$ is an increasing function of $\overline{R}(\mathbf{v})$, the likelihood ratio test is equivalent (in that both tests reject uniformity for large values of $\hat{R}$) to the test described in Section 2.4. (This is a slight extension of the standard result that the likelihood ratio test of uniformity within the von Mises distributions is equivalent to the Rayleigh test; see Section 6.3.1 of Mardia & Jupp, 2000.)

2.4 | Testing quantality

The estimate $\hat{\mathbf{v}}$ can be used as the basis of a test of quantality, that is, that the quantal model (4) fits $\mathbf{u}^\top\mathbf{x}_1, \ldots, \mathbf{u}^\top\mathbf{x}_n$ for some unit vector $\mathbf{u}$ and some quantum q in $(q_{\min}, q_{\max})$. Large values of $\hat{R}$ provide evidence against $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ being a random sample from the uniform distribution on the circle, and so such values support the idea that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ lie near equally spaced parallel lines. Significance of $\hat{R}$ can be assessed by simulation based on the fact that if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ arise as n independent points generated by some Poisson point process then $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ resemble points obtained from the uniform distribution on the circle. In the jth of $B-1$ simulations (for some suitable B), the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are replaced by n independent points in some rectangle containing $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in which the x and y coordinates of the simulated points are independent and uniform on appropriate intervals. For each j the corresponding value $\hat{R}_j$ of $\hat{R}$ is evaluated. It is appropriate to take the p-value as the proportion of $\hat{R}, \hat{R}_1, \ldots, \hat{R}_{B-1}$ that are greater than or equal to the observed $\hat{R}$.

2.5 | Confidence regions for v

It is often appropriate to consider $\mathbf{v}$ as an interest parameter and (μ, κ) as nuisance parameters. We now give two methods of assessing the precision of the estimate $\hat{\mathbf{v}}$ of $\mathbf{v}$.

For the von Mises model (17), approximate confidence regions can be obtained using standard large-sample asymptotics. Because $\overline{R}(\mathbf{v})$ in Equation (13) is a smooth function of $\mathbf{v}$, the profile

log-likelihood (22) for the von Mises model is smooth. Thus, for $0 < \alpha < 1$ and for large n, the deviance regions

$$\left\{ \mathbf{v} : 2[\ell(\hat{\mathbf{v}}) - \ell(\mathbf{v})] \leq \chi^2_{2;100(1-\alpha)} \right\} \tag{26}$$

are approximate $100(1-\alpha)\%$ confidence regions. The coverage probability of the regions (26) can be quite different from $100(1-\alpha)$ unless the region is quite closely clustered around $\hat{\mathbf{v}}$.

For the general model (5), one way of obtaining confidence regions for $\mathbf{v}$ is to compare $\hat{\mathbf{v}}$ with corresponding estimates based on pseudo-data that are obtained by resampling from the residuals, as follows. For a suitably large N_{sims}, put

$$\mathbf{x}_i^{*(j)} = \hat{\mathbf{x}}_i + (\hat{q}/2\pi)\varepsilon_i^{*(j)}\hat{\mathbf{u}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, N_{\text{sims}}, \tag{27}$$

where $\hat{\mathbf{x}}_i$ is the i-th fitted point, defined in Equation (15), and $\varepsilon_1^{*(j)}, \ldots, \varepsilon_n^{*(j)}$ are a sample (with replacement) from the residuals $r_1, \ldots, r_n$. Denote by $\hat{\mathbf{v}}^{(j)}$ the estimate of $\mathbf{v}$ based on $\mathbf{x}_1^{*(j)}, \ldots, \mathbf{x}_n^{*(j)}$.

Recall from the beginning of Section 2 that the axis $\pm\mathbf{u}$ can be represented by the 2×2 matrix $\mathbf{u}\mathbf{u}^\top$. It is only for computational convenience that $\mathbf{u}$ has been restricted to some arbitrary semi-circle. Confidence regions for $\pm\mathbf{v}$ can be obtained using a suitable measure of squared distance between multiples $\pm\mathbf{w}_1$ and $\pm\mathbf{w}_2$ of axes. A convenient such measure is the squared matrix norm of the difference between the corresponding symmetric matrices $\mathbf{w}_1\mathbf{w}_1^\top$ and $\mathbf{w}_2\mathbf{w}_2^\top$, i.e.

$$\begin{aligned} d(\pm\mathbf{w}_1, \pm\mathbf{w}_2) &= \|\mathbf{w}_1\mathbf{w}_1^\top - \mathbf{w}_2\mathbf{w}_2^\top\|_{HS}^2 \\ &= \operatorname{tr}\left[\left(\mathbf{w}_1\mathbf{w}_1^\top - \mathbf{w}_2\mathbf{w}_2^\top\right)^2\right] = \|\mathbf{w}_1\|^4 - 2\left(\mathbf{w}_1^\top\mathbf{w}_2\right)^2 + \|\mathbf{w}_2\|^4, \end{aligned} \tag{28}$$

where $\|\cdot\|_{HS}$ denotes the Hilbert–Schmidt norm (alias the Frobenius norm). For $0 < \alpha < 1$, define c_α as the $\lfloor (1-\alpha)N_{\text{sims}} \rfloor$-th order statistic of $d(\pm\hat{\mathbf{v}}^{(1)}, \pm\hat{\mathbf{v}}), \ldots, d(\pm\hat{\mathbf{v}}^{(N_{\text{sims}})}, \pm\hat{\mathbf{v}})$, where $\lfloor\cdot\rfloor$ denotes the integer part. Then the confidence region

$$\{\pm\mathbf{v} : d(\pm\mathbf{v}, \pm\hat{\mathbf{v}}) < c_\alpha\} \tag{29}$$

for $\pm\mathbf{v}$ has asymptotic coverage $100 \times (1-\alpha)\%$ as $n, N_{\text{sims}} \to \infty$. The corresponding confidence region for $\mathbf{v}$ is obtained by restricting $\mathbf{v}$ to lie in the same half-plane as the semi-circle containing permissible values of $\mathbf{u}$.

If a test of von Misesness does not reject assumption (17) then an alternative to using (27) is to take $\mathbf{x}_1^{*(j)}, \ldots, \mathbf{x}_n^{*(j)}$ (for $j = 1, \ldots, N_{\text{sims}}$) to be a random sample from the fitted von Mises distribution.

Confidence intervals for μ and κ can be obtained from confidence regions for $\mathbf{v}$ by using (12) and (21).

2.6 | Assessing spatial homogeneity

The general model given by hypothesis (5) is spatially homogeneous in that the distribution of $\theta_i(\mathbf{v})$ does not depend on $\mathbf{x}_i$. One way of assessing such spatial homogeneity is by means of a plot (such as a contour plot or heat plot) of the residuals r_i, defined by Equation (14), against the fitted points $\hat{\mathbf{x}}_i$ for $i = 1, \ldots, n$. A pronounced pattern in this plot may indicate heterogeneity. For data

sets of moderate size, a three-dimensional plot of the residuals against the fitted points can be useful.

2.7 | The effect of limited resolution

In general, limitations on the resolution of measurements mean that the measured positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not the (unknown) true positions $\mathbf{z}_1, \ldots, \mathbf{z}_n$ but some perturbed versions of them. Let ξ_1 and ξ_2 be the resolutions in the x-direction and y-direction respectively. Then $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are on a rectangular lattice with distances $2\xi_1$ and $2\xi_2$ between adjacent points in the x-direction and y-direction respectively. For each i, $\mathbf{x}_i$ is the nearest lattice point to $\mathbf{z}_i$.

Define $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ by Equation (7) and let $\hat{R}$ be the maximised mean resultant length of $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$, as in Equation (25). As in Sections 2.3–2.4, it is appropriate to regard large values of $\widehat{R}$ as providing evidence against $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ being a random sample from the uniform distribution on the circle. Thus such values of $\widehat{R}$ support the idea that the unobserved $\mathbf{z}_1, \ldots, \mathbf{z}_n$ lie near equally spaced parallel lines. Significance of $\widehat{R}$ can be assessed by simulation based on the fact that if $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are distributed independently and uniformly on some rectangle with sides parallel to the coordinate axes then $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are distributed independently and uniformly on the corresponding rectangular part of the lattice.

As the awareness that the data are rounded provides no change in the information available, it is appropriate to estimate the parameter $\mathbf{v}$ in the limited-resolution model by $\hat{\mathbf{v}}$ defined in definition (11). Thus the estimate is the same for both the limited-resolution model and the perfect-resolution model of Section 2.1. On the other hand, the significance of the value of $\widehat{R}$ in a test of quantality (i.e. $\theta_1(\mathbf{v}), \ldots, \theta_n(\mathbf{v})$ being a random sample from a non-uniform distribution on the circle) depends on the model; the limited-resolution model is a coarsening of the perfect-resolution model.

2.8 | Simulation study

The performance of the test of quantality, the estimator $\hat{\mathbf{v}}$ given by definition (11), and the confidence region (29), was investigated in a small simulation study.

For $n = 50, 100, 200, 300$ and $\kappa = 0.5, 0.75, 1, 1.5$, N_{sims} independent sets $\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_n^{(j)}$ ($j = 1, \ldots, N_{\text{sims}}$) of n independent points were simulated over a $(0, 125) \times (1.75, 75.25)$ rectangle (similar to that arising in the Catcraig data set). For $i = 1, \ldots, n$ and $j = 1, \ldots, N_{\text{sims}}$, $\mathbf{x}_i^{(j)} = (w_i^{(j)}, 3.5\, z_i^{(j)} + \eta_i^{(j)})$, where $w_i^{(j)}, z_i^{(j)}$ and $\eta_i^{(j)}$ were independent, $w_i^{(j)}$ was distributed uniformly on $(0, 125)$, $z_i^{(j)}$ was distributed uniformly on $\{1, \ldots, 21\}$, and $(2\pi/3.5)\eta_i^{(j)}$ had the $M(-2.57, \kappa)$ distribution.

For each n and κ, let Δ denote the proportion of the N_{sims} simulated samples for which quantality is detected at the 5% level, using the test of Section 2.4 that compares the value of $\widehat{R}$ with values obtained from $B - 1$ samples of n uniformly distributed points.

We define the *(empirical) root mean square error* of $\pm\hat{\mathbf{v}}$ as an estimator of $\pm\mathbf{v}_0$ by

$$\text{rmse} = \left[\frac{1}{N_{\text{sims}}} \sum_{j=1}^{N_{\text{sims}}} d\left(\pm\hat{\mathbf{v}}^{(j)}, \pm\mathbf{v}_0 \right) \right]^{1/2}, \tag{30}$$

TABLE 1 Performance of test of quantality, estimators and 95% confidence regions

n	κ	Δ (%)	$(\overline{u}, \overline{q})$	rmse	cov (%)
50	0.50	8	(84.8, 2.68)	8.07	88
	0.75	11	(88.2, 2.79)	6.60	87
	1.00	36	(88.8, 3.08)	3.40	90
	1.50	93	(89.6, 3.50)	0.30	93
100	0.50	16	(84.2, 2.91)	6.56	87
	0.75	59	(91.1, 3.29)	3.29	92
	1.00	94	(89.4, 3.50)	0.57	93
	1.50	100	(90.0, 3.50)	0.02	91
200	0.50	50	(85.4, 3.27)	4.15	92
	0.75	98	(89.2, 3.49)	1.06	91
	1.00	100	(90.0, 3.50)	0.02	92
	1.50	100	(90.0, 3.50)	0.01	94
300	0.50	82	(84.2, 2.91)	2.37	91
	0.75	100	(91.1, 3.29)	0.02	90
	1.00	100	(89.4, 3.50)	0.02	92
	1.50	100	(90.0, 3.50)	0.01	92

Based on $N_{\text{sims}} = 100$ ($N_{\text{sims}} = 500$ for cov) simulations from artificial data sets described in text. Δ is the proportion of the N_{sims} simulated samples for which quantality is detected at the 5% level by the test of Section 2.4 with $B = 2000$. $(\overline{u}, \overline{q})$ is the mean of estimates $(\hat{u}^{(1)}, \hat{q}^{(1)}), \ldots, (\hat{u}^{(N_{\text{sims}})} \hat{q}^{(N_{\text{sims}})})$. rmse is the empirical root mean square error defined in (30). cov is the empirical coverage probability of the 95% confidence regions (29).

where, for $j = 1, \ldots, N_{\text{sims}}$, $\hat{\mathbf{v}}^{(j)}$ is the estimate of $\mathbf{v}$ based on $\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_n^{(j)}$ given by definition (11) and d is defined in (28).

The performances of the test of quantality given in Section 2.4, the estimators $\hat{\mathbf{v}}$, and the 95% confidence regions (29) are summarised in Table 1. The value of $\mathbf{v}_0$ was $(2\pi/3.5)(0, 1)^\top$. The values of Δ indicate that, not surprisingly, the test of quantality performs poorly when $\kappa = 0.50$ (so that the samples tend to be nearly uniform), especially for small n, but its performance improves markedly as κ increases, with the probability of detecting quantality for $\kappa = 1.50$ being very high even for moderately small n. As n and κ increase, rmse decreases, and so the accuracy of $\pm\hat{\mathbf{v}}$ as an estimator of $\pm\mathbf{v}_0$ increases. Table 1 shows also that, while the coverage of the 95% confidence regions falls below the nominal 95% level, this coverage is above 90%, except when both n and κ are small.

`R` code for implementation of the techniques described in this Section is available at http://www.mcs.st-and.ac.uk/~pej/quantal. With the chosen default values of the program control parameters, for the Catcraig data, running on a 2021 MacBook Pro laptop (10-core M1 Pro processor with 32 GB memory), the first 10 sections of the code run in less than 8 s, while the combined running time of the tests of quantality in sections 11 and 12 is around 80 s and the confidence regions in section 13 are produced in around 410 s.

3 | DATA ANALYSIS

In this section we analyse the Cellardyke and Catcraig data sets, as described in Section 1.

3.1 | Cellardyke data set

The Cellardyke data set shown in Figure 1 was analysed using the methods of Section 2. The p-value of the test of uniformity (given in Section 2.4) based on $B = 1{,}000$ simulations was 0.18, and so we deduce that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ do not lie near equally spaced parallel lines. The same conclusion is obtained from the limited-resolution model of Section 2.7, for which the p-value of the corresponding test of uniformity is 0.16. Ignoring this lack of fit and nonetheless estimating $\mathbf{v}$ by definition (11) (with q restricted to the range (0.5, 2.5)) leads to the estimate $(\hat{u}, \hat{q}) = (-65°, 1.80)$, where $\hat{\mathbf{u}} = (\cos\hat{u}, \sin\hat{u})^\top$, giving the fitted lines shown in Figure 5. The apparently good fit of model (4) suggested by the Figure is shown by the test not to be significant.

3.2 | Catcraig data set

The Catcraig data set shown in Figure 2 was analysed using the methods of Section 2. The quantum parameter, q, was restricted to the range (1.5, 4.5). The p-value of the test of uniformity based on $B = 1{,}000$ simulations was 0.03, and so we deduce that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ do lie near equally spaced parallel lines.

The heat map of the deviance, $2\{\ell(\hat{\mathbf{v}}) - \ell(\mathbf{v})\}$, in Figure 6 shows that it has a clear global maximum at $(\hat{v}_1, \hat{v}_2) = (1.63, 0.66)$, corresponding to $(\hat{u}, \hat{q}) = (22°, 3.57)$. The fitted lines are shown in Figure 7.

Due to limitations in measuring the centres of saucer-shaped depressions from an aerial image using a ground-based 2 m graduated (10 cm) scale bar, we have $\xi_1 = \xi_2 = 5$ cm. However, Watson's U^2 test is not significant at the 10% level, indicating good fit of $\theta_1(\hat{\mathbf{v}}), \ldots, \theta_n(\hat{\mathbf{v}})$ to a von Mises distribution, so that we can assume that (17) holds. Thus there is no need to use the limited-resolution model of Section 2.7.

Figure 8 is a heat map of the residuals. Since there is no obvious pattern, we have no reason to doubt that the spatial homogeneity implicit in model (5) is appropriate here.

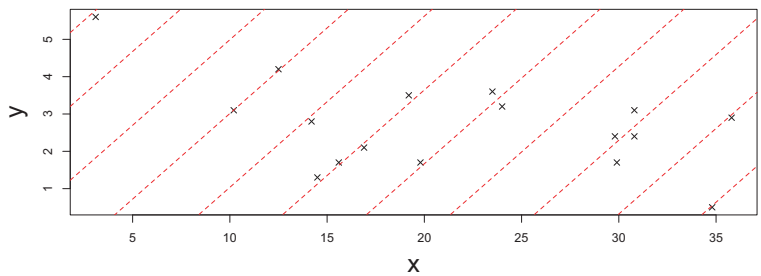

FIGURE 5 Equally spaced parallel lines fitted to Cellardyke data set

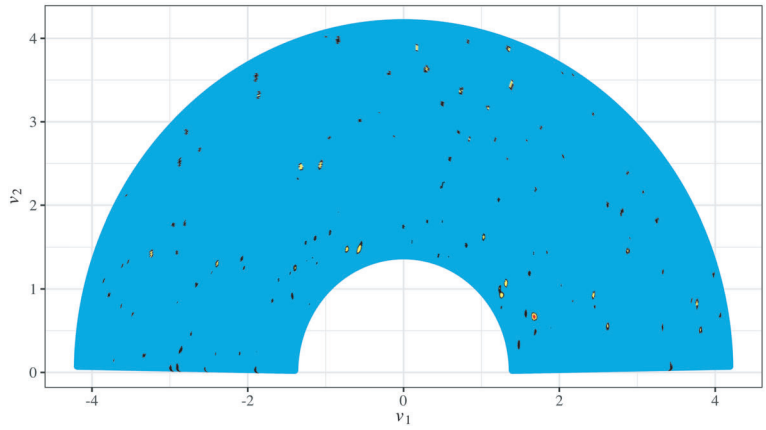

FIGURE 6 Heat map of deviance, $2\{\ell(\hat{\mathbf{v}}) - \ell(\mathbf{v})\}$, as function of $\mathbf{v}$ for Catcraig data set. The 95% deviance region for $\mathbf{v}$ is indicated by the red region. Corresponding regions at the 99.9% and 99.99% levels are those obtained by successive additional inclusion of the yellow and black regions respectively

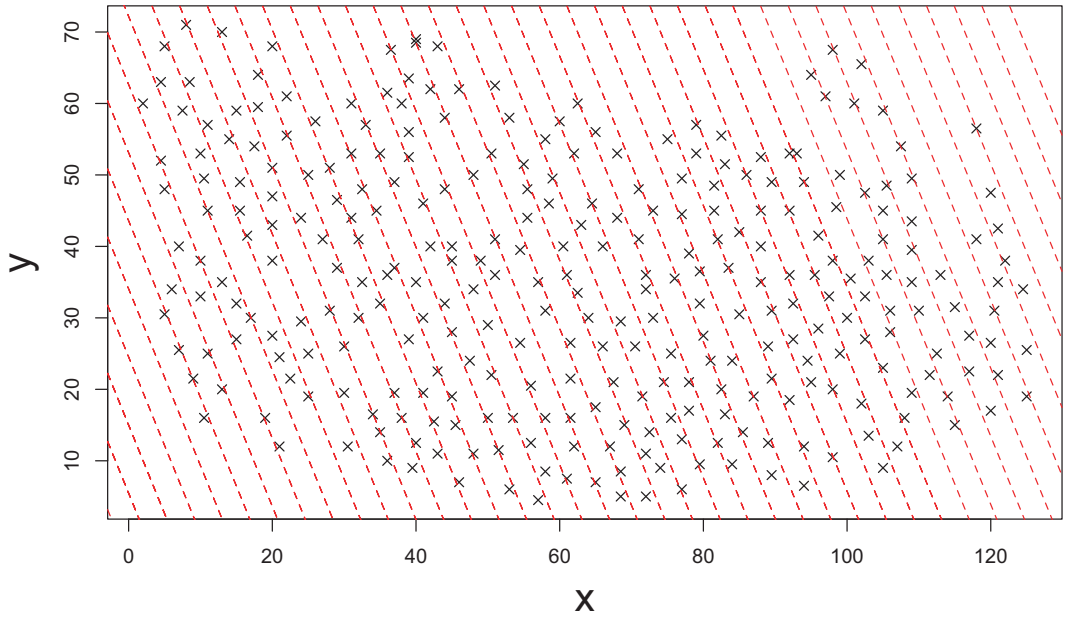

FIGURE 7 Equally spaced parallel lines fitted to Catcraig data set

The spikiness of the mean resultant length in Figure 4 suggests that the deviance regions (26) will not have the nominal coverage probability of $100(1-\alpha)\%$ unless the region is quite closely clustered around $\hat{\mathbf{v}}$. The 95% deviance region $\{\mathbf{v} : 2[\ell(\hat{\mathbf{v}}) - \ell(\mathbf{v})] \leq \chi^2_{2;0.95}\}$ (shown in red in Figure 6) is tightly clustered round $\hat{\mathbf{v}}$ and is almost elliptical. On the other hand, the 99.9% deviance region $\{(\mathbf{v} : 2[\ell(\hat{\mathbf{v}}) - \ell(\mathbf{v})] \leq \chi^2_{2;0.999}\}$ (shown in red and yellow) is not even connected.

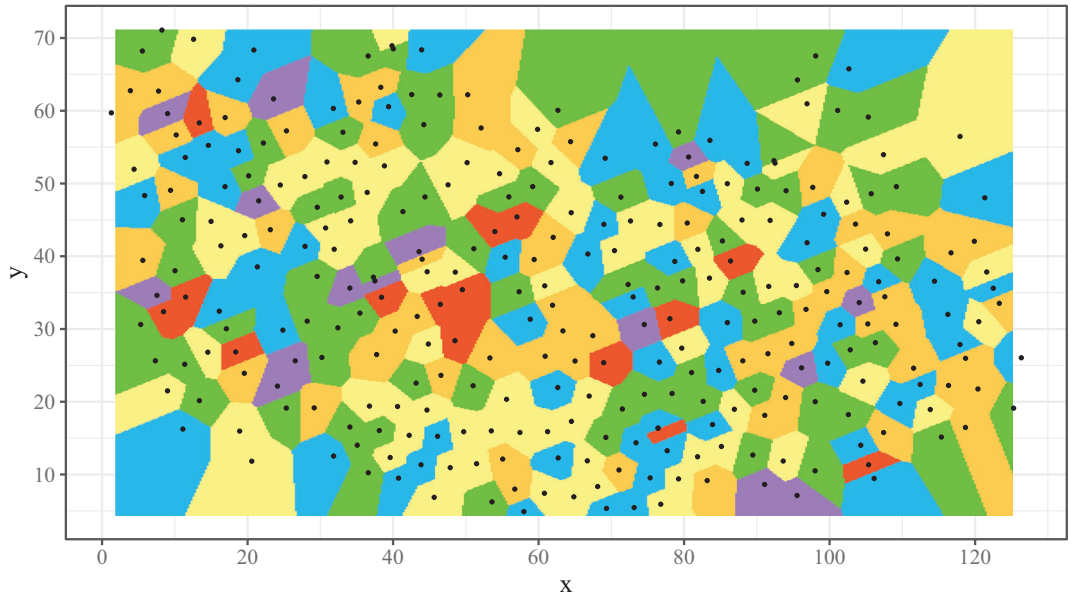

FIGURE 8 Heat map of the residuals, taking values in $[-\pi, \pi)$, for the Catcraig data set. The areas coloured red, orange, yellow, green, blue and purple are separated by contours at 2.76, 1.08, −0.021, −1.15 and −2.86, respectively, corresponding to the 95th, 75th, 50th, 25th and 5th percentiles of the residuals. Black dots denote fitted points. Units of measurement of axes in metres

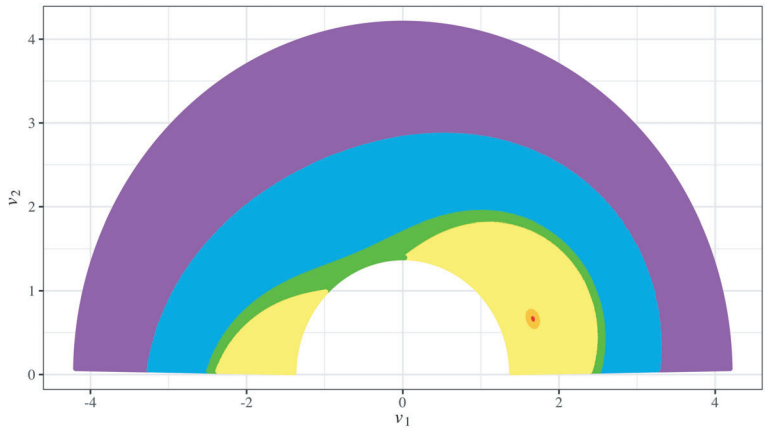

FIGURE 9 80% confidence region for $\mathbf{v}$ (shown in red) given by Equation (29) with $N_{sims} = 25{,}600$ for the Catcraig data set. Corresponding regions at the 82%, 84%, 90% and 95% levels are those obtained by successive additional inclusion of orange, yellow, green and blue areas respectively

Figure 9 shows 80%, 82%, 84%, 90% and 95% confidence regions (29) for $\mathbf{v}$ obtained from pseudo-data generated by resampling from the residuals, as described in Section 2.5. These non-parametric confidence regions are much larger than the deviance regions of Figure 6. They are also much smoother as, unlike the deviance regions, they do not reflect the spikiness of the underlying likelihood surface.

4 | DISCUSSION

A natural three-dimensional analogue of conjecture (1) is the conjecture for given points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in 3-space that

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \text{ lie 'near' equally spaced parallel planes in 3-space.} \tag{31}$$

Then conjecture (31) can be handled using (6)–(12), where now $\mathbf{u}$ is a unit vector in 3-space. The 3-vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are transformed to the real numbers $\mathbf{u}^\top\mathbf{x}_1, \ldots, \mathbf{u}^\top\mathbf{x}_n$ and then to the angles $2\pi\{\mathbf{u}^\top\mathbf{x}_1/q\}, \ldots, 2\pi\{\mathbf{u}^\top\mathbf{x}_n/q\}$. Equally spaced parallel planes occur as very narrow planes of glassy material in grains of silicate materials; see https://en.wikipedia.org/wiki/Planar_deformation_features or Langenhorst and Deutsch (1993).

It is straightforward to extend the methods of this paper to fit (a) several sets of equally spaced parallel lines, (b) several sets of equally spaced concentric circles, to points in the plane. If the normals to any two sets of equally spaced parallel lines are nearly parallel then it is difficult to estimate their directions. Similarly, if any two sets of equally spaced concentric circles are almost concentric then it is difficult to estimate their centres. Furthermore, unless the observed points near a circle lie on a large arc, it is not possible to estimate the centre very precisely.

ACKNOWLEDGEMENTS

R. J. B. Goudie was funded by the UKRI Medical Research Council (MRC) [programme code MC_UU_00002/2] and supported by the NIHR Cambridge Biomedical Research Centre. We are grateful to the Associate Editor and two referees for their helpful comments on an earlier version of this manuscript. These have led to considerable improvements.

DATA AVAILABILITY STATEMENT

The datasets (and our R code) are available at MailScanner has detected a possible fraud attempt from "urldefense.com" claiming to be https://www.mcs.st-and.ac.uk/~pej/quantal.

ORCID

P.E. Jupp https://orcid.org/0000-0003-0973-8434
I.B.J. Goudie https://orcid.org/0000-0002-3910-7310
R.J.B. Goudie https://orcid.org/0000-0001-9554-1499

REFERENCES

Batchelor, R.A., Garton, R.E. & Jupp, P.E. (submitted) Seismites in Carboniferous sediments, Dunbar, Scotland.

Çankaya, E. & Fieller, N.R.J. (2009) Quantal models: a review with additional methodological development. *Journal of Applied Statistics*, 36, 369–384.

Fisher, N.I. (1993) *Statistical analysis of circular data*. Cambridge: Cambridge University Press.

Freeman, P.R. (1976) A Bayesian analysis of the megalithic yard. *Journal of the Royal Statistical Society: Series A*, 139, 20–35.

Jammalamadaka, S.R. & SenGupta, A. (2001) *Topics in circular statistics*. Singapore: World Scientific.

Kendall, D.G. (1974) Hunting quanta. *Philosophical Transactions of the Royal Society A*, 76, 231–266.

Langenhorst, F. & Deutsch, A. (1993) Orientation of planar deformation features (PDFs) in quartz. *Abstracts of 24th Lunar and Planetary Science Conference, Houston, Texas, 15-19 March 1993*, 849.

Mardia, K.V. & Jupp, P.E. (2000) *Directional statistics*. Chichester: Wiley.

von Mises, R. (1918) Über die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift*, 19, 490–500.

Pakkanen, J. (2002) Deriving ancient foot units from building dimensions: a statistical approach employing cosine quantogram analysis. In Burenhult, G. & Arvidsson, J. (Eds.) *Archaeological informatics: pushing the envelope. CAA 2001*, Oxford: Archaeopress. pp. 501–506.

Pewsey, A., Neuhäuser, M. & Ruxton, G.D. (2013) *Circular statistics in R*. Oxford: Oxford University Press.

Resor, P.G. & Meer, V.E. (2009) Slip heterogeneity on a corrugated fault. *Earth and Planetary Science Letters*, 288, 483–491.

How to cite this article: Jupp, P.E., Goudie, I.B.J., Batchelor, R.A. & Goudie, R.J.B. (2022) Assessing seismic origin of geological features by fitting equidistant parallel lines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–16. Available from: https://doi.org/10.1111/rssc.12553