# CURRENT TREND AND STUDIES IN NATURAL SCIENCE AND MATHEMATICS

All Sciences Academy

# *CURRENT TREND AND STUDIES IN NATURAL SCIENCE AND MATHEMATICS*

**Editor**
**Prof. Dr. Canan ÖZDEMİR**

***Current Trend and Studies in Natural Science and Mathematics***
*Editor: Prof. Dr. Canan ÖZDEMİR*

**Design:** All Sciences Academy Design
**Published Date:** September 2026
**Publisher's Certification Number:** 72273
**ISBN:** 978-625-5472-01-4



www.allsciencesacademy.com
allsciencesacademy@gmail.com

## CONTENT

# Stability Analysis and Numerical Investigation of Turing Pattern Formation in a Schnakenberg Reaction–Diffusion System

**Özlem KAYTMAZ**

Assoc. Prof. Dr., Zonguldak Bülent Ecevit University, Faculty of Science, Department of Mathematics, Division of Applied Mathematics, Türkiye ozlem.kaytmaz@beun.edu.tr ORCID No: 0000-0003-0420-007X

## ABSTRACT

This chapter considers Turing pattern formation in a two-component Schnakenberg reaction–diffusion system by combining linear stability analysis with numerical computation. The positive homogeneous equilibrium, the corresponding Jacobian matrix, and the conditions for local kinetic stability are first derived. Modal analysis of the linearized reaction–diffusion system then gives the dispersion relation, the conditions for diffusion-driven instability, and the unstable wavenumber interval. The predicted unstable wavenumber interval is compared with the spatial modes obtained numerically on a finite one-dimensional domain with homogeneous Neumann boundary conditions. The numerical investigation combines second-order finite differences in space with BDF time integration. Controlled single-mode perturbations are used for threshold tests, while a broadband deterministic perturbation allows mode selection to be examined without imposing the theoretically fastest-growing mode. Pattern amplitude and the dominant cosine mode quantify the numerical response, while additional computations examine sensitivity to the activator diffusion coefficient, mesh resolution, and the BDF maximum step. The finite-domain comparison also accounts explicitly for the discrete Neumann spectrum. The computations distinguish kinetic stability from diffusion-driven modal instability and compare the continuous dispersion prediction with the admissible modes of the finite domain. Together, the analytical and numerical results clarify the relation between Turing instability predicted by linear theory and the spatial modes observed in finite-domain computations.

*Keywords – Schnakenberg reaction–diffusion system, Turing instability, pattern formation, stability analysis, dispersion relation, finite differences.*

## INTRODUCTION

Reaction–diffusion equations couple local reaction kinetics with spatial diffusion, allowing the stability of homogeneous states to depend on both reaction and transport parameters. Under suitable conditions, unequal diffusion can destabilize a spatially homogeneous equilibrium that is stable for the corresponding kinetic ordinary differential equation. This mechanism is known as diffusion-driven, or Turing, instability.

Turing (1952:46) showed that the interaction of chemical reactions and diffusion can amplify small departures from an initially nearly homogeneous state and lead to spatial pattern formation. Mathematically, the homogeneous equilibrium remains stable to spatially homogeneous perturbations but becomes unstable to perturbations with nonzero wavenumber.

Linearization about a homogeneous equilibrium separates the effects of reaction kinetics and diffusion: the reaction terms determine the local

Jacobian, while diffusion modifies the growth rate of each spatial mode according to its wavenumber. Maini, Painter and Chau (1997:3601) examined reaction–diffusion mechanisms of spatial pattern formation in chemical and biological systems.

The reaction kinetics considered here are based on the two-component model introduced by Schnakenberg (1979:398), with a relabelling of the constant parameters. Coupling this nonlinear kinetics with diffusion gives a two-component reaction–diffusion system in which diffusion-driven instability and spatial pattern formation can be analysed. For this system, the homogeneous equilibrium, kinetic Jacobian, Turing instability conditions, and unstable wavenumber interval can all be obtained explicitly. The formulation used below contains two positive kinetic parameters, nonlinear reaction coupling, and two positive diffusion coefficients.

Kinetic stability must be distinguished from diffusion-driven instability, since the appearance of spatial structure in a numerical solution alone does not establish a Turing mechanism. The reaction-only equilibrium must first satisfy the local asymptotic stability conditions. After diffusion is introduced, at least one admissible nonzero spatial mode must have a growth rate with positive real part. Accordingly, the unstable wavenumber interval is determined before the computed spatial patterns are interpreted.

The computed patterns may depend on spatial resolution, time integration, domain size, and the perturbation used to break spatial homogeneity. Pattern amplitude and the dominant cosine mode are used to quantify the computed patterns, while mesh refinement and BDF maximum-step tests assess their sensitivity to the numerical discretization.

Wei and Winter (2008:53) studied the existence and stability of stationary multiple-spot solutions for reaction–diffusion systems, with their main analysis presented in the context of the Schnakenberg model. Khan, Ali, Hamadneh, Abdullah and Alam (2021:7) presented a numerical investigation of a fractional-order Schnakenberg model using the Laplace Adomian decomposition method. Woolley, Krause and Gaffney (2021:1) developed a constructive framework for identifying reaction–diffusion systems that exhibit Turing instability within a prescribed parameter region. The present chapter does not introduce a new reaction model or instability mechanism; instead, it combines stability analysis with numerical experiments for a finite-domain Schnakenberg system.

First, the positive homogeneous equilibrium and the conditions for its kinetic stability are derived. Second, the diffusion-modified characteristic equation is obtained for a generic spatial wavenumber and used to determine the unstable wavenumber interval and linear modal growth rates. The resulting instability thresholds are then used to define the numerical experiments. A single-mode perturbation is used to probe growth or decay across selected values of $D_v$, while a broadband deterministic perturbation is used to examine mode selection without prescribing the theoretically fastest-growing mode.

Finally, mesh refinement, BDF maximum-step, domain-size, and perturbation-amplitude tests are used to assess the sensitivity of the numerical results.

This separates three questions: whether the homogeneous kinetic equilibrium is stable, whether diffusion creates an unstable wavenumber interval, and whether the dominant mode observed numerically agrees with the finite-domain spectral prediction. This distinction is especially important near the instability threshold, where small growth or decay rates can produce long transients.

## MATHEMATICAL FORMULATION OF THE SCHNAKENBERG REACTION–DIFFUSION SYSTEM

Let $\Omega = (0, L)$ be a one-dimensional spatial domain and let $u(x,t)$ and $v(x,t)$ denote two interacting concentrations. A standard nondimensional Schnakenberg reaction–diffusion system is written as

$$\frac{\partial u}{\partial t} = D_u \Delta u + \gamma(a - u + u^2 v), \qquad x \in \Omega,\ t > 0, \tag{1}$$

$$\frac{\partial v}{\partial t} = D_v \Delta v + \gamma(b - u^2 v), \qquad x \in \Omega,\ t > 0. \tag{2}$$

where $a > 0$ and $b > 0$ are kinetic parameters, $D_u > 0$ and $D_v > 0$ are constant diffusion coefficients, and $\gamma > 0$ scales the reaction kinetics relative to diffusion. Homogeneous Neumann boundary conditions are imposed,

$$\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0, \qquad x \in \partial\Omega,\ t > 0 \tag{3}$$

together with nonnegative initial data chosen as small spatial perturbations of the homogeneous equilibrium. For the constant diffusion coefficients used here, the homogeneous Neumann conditions impose zero normal diffusive flux at the boundary and yield cosine Laplacian eigenmodes on the one-dimensional interval.

### *Homogeneous equilibrium*

Suppressing diffusion gives the kinetic system $u_t = \gamma f(u,v)$, $v_t = \gamma g(u,v)$, where $f(u,v) = a - u + u^2 v$ and $g(u,v) = b - u^2 v$. A homogeneous steady state $(u^\star, v^\star)$ satisfies $f(u^\star, v^\star) = g(u^\star, v^\star) = 0$. From $g = 0$ one obtains $(u^\star)^2 v^\star = b$. Substitution into $f = 0$ gives $u^\star = a + b$, and therefore

$$u^\star = a + b, \qquad v^\star = \frac{b}{(a+b)^2}. \tag{4}$$

Since $a$ and $b$ are positive, the equilibrium is positive and unique within the positive quadrant for this kinetic form.

### *Linear stability without diffusion*

The Jacobian matrix of the reaction vector $(f, g)$ is

$$J(u,v) = \begin{pmatrix} -1+2uv & u^2 \\ -2uv & -u^2 \end{pmatrix}. \tag{5}$$

At the homogeneous equilibrium, this becomes

$$J^\star = \begin{pmatrix} \frac{b-a}{a+b} & (a+b)^2 \\ -\frac{2b}{a+b} & -(a+b)^2 \end{pmatrix}. \tag{6}$$

Its determinant simplifies to $\det(J^\star) = (a+b)^2 > 0$, whereas its trace is $\mathrm{tr}(J^\star) = (b-a)/(a+b) - (a+b)^2$. Because $\gamma$ is positive, scaling the kinetic Jacobian by this factor preserves the signs required by the two-dimensional trace-determinant stability test. Thus, the reaction-only equilibrium is locally asymptotically stable when

$$\mathrm{tr}(J^\star) < 0, \qquad \det(J^\star) > 0. \tag{7}$$

For the Schnakenberg kinetics the determinant condition is automatically satisfied for positive $a$ and $b$, so kinetic stability is controlled by the trace inequality.

## TURING INSTABILITY AND THE DISPERSION RELATION

To determine whether diffusion destabilizes the stable homogeneous equilibrium, write $u = u^\star + U$ and $v = v^\star + V$ and retain only linear terms. For a Laplacian eigenmode with $-\Delta\varphi = k^2\varphi$, the modal amplitudes satisfy a two-dimensional linear system with matrix $M(k) = \gamma J^\star - k^2 D$, where $D = \mathrm{diag}(D_u, D_v)$. The growth rates $\lambda = \lambda(k)$ therefore satisfy

$$\lambda^2 - T(k^2)\lambda + H(k^2) = 0, \tag{8}$$

where

$$T(q) = \gamma\,\mathrm{tr}(J^\star) - (D_u + D_v)q, \qquad q = k^2, \tag{9}$$

$$H(q) = D_u D_v q^2 - \gamma(D_v f_u + D_u g_v)q + \gamma^2 \det(J^\star). \tag{10}$$

If the kinetic equilibrium is stable, then $T(q) < 0$ for every $q \geq 0$. Hence a positive modal growth rate occurs precisely when $H(q) < 0$ for some $q > 0$. Since $H$ is an upward-opening quadratic in $q$, diffusion-driven instability requires

$$D_v f_u + D_u g_v > 0, \tag{11}$$

$$(D_v f_u + D_u g_v)^2 - 4 D_u D_v \det(J^\star) > 0. \tag{12}$$

Conditions (7), (11), and (12), together with positive diffusion coefficients, give the two-species Turing instability criterion in the present notation. They separate stability of the homogeneous kinetic mode from instability of nonzero spatial modes caused by diffusion.

### *Unstable wavenumber interval*

When the discriminant in (12) is positive, the two roots of $H(q) = 0$ are

$$q_{\pm} = \frac{\gamma\left[(D_v f_u + D_u g_v) \pm \sqrt{(D_v f_u + D_u g_v)^2 - 4D_u D_v \det(J^\star)}\right]}{2D_u D_v}. \quad (13)$$

Modes satisfying $q_- < k^2 < q_+$ have $H(k^2) < 0$ and are linearly unstable. Under homogeneous Neumann conditions on $\Omega = (0, L)$, the admissible wavenumbers are $k_n = n\pi/L$, $n = 0,1,2, \ldots$ . Thus, a continuous unstable band does not imply finite-domain linear instability unless at least one admissible $k_n$ lies inside the interval. This domain-spectrum condition is checked explicitly in the numerical experiments.

***Reference parameter set***

For $a = 0.1$, $b = 0.9$, $D_u = 1$, $D_v = 10$, and $\gamma = 1$, the equilibrium is $(u^\star, v^\star) = (1,0.9)$. At this point $f_u = 0.8$, $f_v = 1$, $g_u = -1.8$, and $g_v = -1$. The trace is $-0.2$ and the determinant is 1, so the kinetic equilibrium is stable. Moreover, $D_v f_u + D_u g_v = 7 > 0$ and the Turing discriminant equals $49 - 40 = 9 > 0$. Equation (13) then gives $q_- = 0.2$ and $q_+ = 0.5$. For the computational domain $L = 20\pi$ used below, the admissible Neumann wavenumbers are $k_n = n\pi/L = n/20$. Hence modes $n = 9, \ldots, 14$ lie in the unstable band, providing a direct finite-domain test of the dispersion relation.

## NUMERICAL DISCRETIZATION AND COMPUTATIONAL DIAGNOSTICS

A uniform grid $x_i = ih, i = 0, \ldots, N$, with $h = L/N$ is used. The Laplacian is approximated by the centered second-order difference in the interior, with homogeneous Neumann conditions imposed by an even-reflection endpoint completion. The resulting semidiscrete nonlinear system is advanced in time by a variable-order backward differentiation formula (BDF) solver. The method-of-lines formulation is suitable for the resulting stiff semidiscrete system, while implicit BDF integration avoids the severe diffusion-driven stability restriction associated with a fully explicit time discretization. All numerical values reported below are generated from this implementation. The resulting semidiscrete system can be written as

$$\Delta_h w_i = \frac{w_{i-1} - 2w_i + w_{i+1}}{h^2}. \quad (14)$$

The initial data are constructed as small deterministic perturbations of $(u^\star, v^\star)$. Using the same controlled perturbation across parameter sets permits direct comparison of growth and decay. The primary scalar diagnostic is the peak-to-peak pattern amplitude

$$A_u(t) = \max_i u_i(t) - \min_i u_i(t). \quad (15)$$

An analogous peak-to-peak amplitude is used for the inhibitor variable. Modal content is estimated using a type-I discrete cosine transform (DCT-I), consistent with the endpoint-inclusive grid and the Neumann cosine

eigenfunctions. The dominant numerical mode is then compared with the admissible modes predicted by the dispersion relation.

***Numerical experiment protocol***

The computations are organized to separate threshold behavior from wavelength selection. First, a single admissible cosine perturbation is used as a controlled probe below, near, and above the analytically computed Turing threshold; this experiment tests decay or growth of a prescribed unstable candidate mode and is not used as evidence of spontaneous mode selection. Second, a separate broadband deterministic perturbation containing cosine modes $n = 1, \dots, 20$ is used for the reference case so that the dominant numerical mode can emerge from competition among several admissible components. Because all modes $n = 1, \dots, 20$ enter this perturbation with equal modal weights, the dominant $n = 11$ mode observed later is not prescribed by the initial condition. Finally, mesh refinement, BDF maximum-step sensitivity, variation in $D_v$, perturbation-amplitude tests, and finite-domain experiments are used to assess robustness.

## *TIME INTEGRATION AND DISCRETE NEUMANN OPERATOR*

For the computations reported below, the method-of-lines formulation is used. The second-order centered finite-difference approximation is retained in space, while the resulting stiff nonlinear ODE system is integrated by a variable-order backward differentiation formula (BDF) solver with relative tolerance $10^{-7}$ and absolute tolerance $10^{-9}$. The purpose of this choice is to resolve the growth and saturation of diffusion-driven modes in the coupled semilinear system without imposing the restrictive diffusion step-size condition associated with a fully explicit time discretization.

At the two endpoints, the homogeneous Neumann condition is imposed by even reflection. Thus, if $U_{\mathrm{i}}$ approximates $u(x_i, t)$, the centered second-order Laplacian is used at interior nodes, while the endpoint values are completed by the even-reflection formulas displayed in (16). The same construction is applied to $v$. These endpoint completions are second-order consistent with an even extension of a sufficiently smooth Neumann solution.

$$(\Delta_h U)_0 = \frac{2(U_1 - U_0)}{h^2}, \qquad (\Delta_h U)_N = \frac{2(U_{N-1} - U_N)}{h^2}. \tag{16}$$

The domain length is $L = 20\pi$. For homogeneous Neumann conditions, the admissible wavenumbers are $k_n = n\pi/L = n/20$. For the reference parameters $D_u = 1$ and $D_v = 10$, the unstable interval is $0.4472136 < k < 0.7071068$. Consequently, the admissible unstable indices satisfy approximately $n = 9, \dots, 14$, so modes $n = 9, \dots, 14$ are predicted to be unstable. Direct maximization of the larger eigenvalue of $\gamma J^\star - k^2 \mathrm{diag}(D_u, D_v)$ gives the fastest-growing continuous wavenumber $k_{\mathrm{pred}} \approx 0.5644$. Among the admissible unstable Neumann modes, the largest discrete

linear growth rate occurs at $n = 11$ ($k_{11} = 0.55$). This provides a quantitative prediction for independent testing in the nonlinear simulation.

## COMPUTATIONAL EXPERIMENTS

### *Threshold calculation and parameter classification*

For fixed $a = 0.1$, $b = 0.9$, and $D_u = 1$, the kinetic Jacobian is fixed. The continuous-wavenumber Turing discriminant is therefore a quadratic function of $D_v$. Its two algebraic zeros are approximately $0.18237$ and $8.56763$; together with the positivity condition, the relevant continuous threshold is $D_{\mathrm{v,c}}^{cont} \approx 8.56763$. On the finite Neumann domain $L = 20\pi$, however, only $k_n = n\pi/L = n/20$ are admissible. The first admissible mode to lose stability is $n = 12$ at $D_v \approx 8.58586$, whereas the controlled probe $n = 11$ crosses at $D_v \approx 8.65484$. Thus $D_v = 8.6$ is above the continuous and finite-domain onset, but the particular $n = 11$ probe remains linearly stable. This discrete-spectrum distinction is important when interpreting the near-threshold experiment. Accordingly, the critical diffusion threshold is given by

$$D_{v,c} \approx 8.56763. \tag{17}$$

### *Amplitude evolution across the Turing threshold*

For the threshold-oriented experiment, all runs use the controlled single-mode perturbations $u(x,0) = u^\star + 0.01\cos(11\pi x/L)$ and $v(x,0) = v^\star + 0.005\cos(11\pi x/L)$. Because mode 11 is prescribed, this experiment measures the behavior of that particular spatial component and is not used to infer spontaneous wavelength selection. Table 1 reports $A_u(t)$. For $D_v = 8$ the perturbation decays strongly. At $D_v = 8.6$ the amplitude shows a short transient increase and then decays; this is consistent with the finite-domain calculation because mode 11 does not cross its own linear threshold until $D_v \approx 8.65484$. For $D_v = 9,10,15$, mode 11 is unstable and sustained growth is observed, followed by nonlinear saturation in the more strongly unstable cases.

Table 1. Activator amplitude across the diffusion-driven threshold ($N = 128$).

| $D_v$ | $A_u(0)$ | $A_u(25)$ | $A_u(50)$ | $A_u(100)$ | $A_u(200)$ |
|---|---|---|---|---|---|
| 8 | 0.020000 | 0.011460 | 0.004814 | 0.000850 | 0.000026 |
| 8.6 | 0.020000 | 0.024514 | 0.022674 | 0.019403 | 0.014223 |
| 9 | 0.020000 | 0.038526 | 0.056810 | 0.121388 | 0.379565 |
| 10 | 0.020000 | 0.102347 | 0.378966 | 0.861886 | 0.868642 |
| 15 | 0.020000 | 1.196481 | 1.466079 | 1.466251 | 1.466251 |

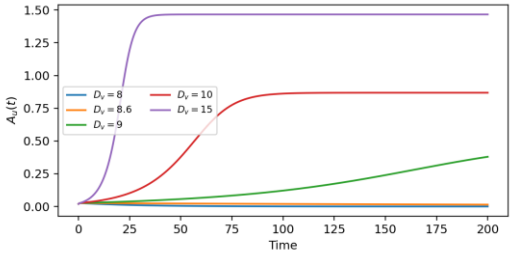



Figure 1. Evolution of the activator amplitude for selected inhibitor diffusion coefficients.

***Predicted and observed dominant mode***

To test wavelength selection independently of the prescribed mode used in the threshold experiment, the reference case $D_v = 10$ was repeated from the broadband deterministic perturbation defined in (18), which contains equal small cosine components with indices $n = 1, \dots, 20$ in both variables (with the $v$ perturbation scaled by one half). The continuous dispersion relation attains its maximum growth rate at $k_{\text{pred}} \approx 0.5644$. On $L = 20\pi$, the admissible Neumann wavenumbers are $k_n = n\pi/L = n/20$, and among the unstable modes $n = 9, \dots, 14$, the largest discrete growth rate occurs for $n = 11$, corresponding to $k_{11} = 0.55$. Starting from the broadband perturbation, a type-I discrete cosine transform (DCT-I), compatible with the endpoint-inclusive grid and the cosine eigenmodes associated with the homogeneous Neumann boundary conditions, identifies $n = 11$ as the dominant mode at the final time. This separate experiment therefore allows the numerically emerging dominant mode to be compared directly with the pre-computed finite-domain spectral prediction. For this experiment, the deterministic perturbation and the corresponding initial data are defined by

$$p(x) = \frac{1}{20}\sum_{n=1}^{20} cos(n\pi x/L),$$
$$u(x,0) = u^\star + 0.01p(x), \quad v(x,0) = v^\star + 0.005p(x). \qquad (18)$$

For comparison, Figure 2 shows the temporal evolution of the activator profile for the controlled single-mode reference case at $D_v = 10$ and $N = 256$; it is distinct from the broadband mode-selection experiment described above.

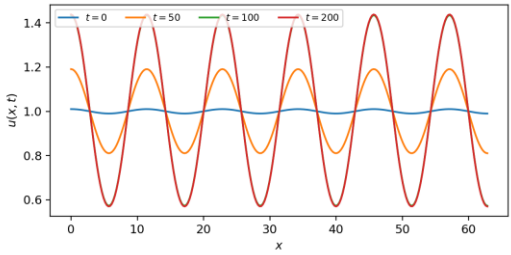



Figure 2. Activator profiles at selected times for the controlled single-mode reference case $D_v = 10$ on $L = 20\pi$ ($N = 256$).

### *Mesh refinement study*

A mesh study was carried out for $D_v = 10$ with $N = 64, 128, 256$ using the broadband deterministic perturbation, while keeping the physical domain and solver tolerances fixed. The final pattern amplitude and dominant cosine mode are reported in Table 2. The dominant index remains $n = 11$ on all three meshes. The final amplitude changes only slightly under refinement from $N = 128$ to $N = 256$. The results show weak sensitivity to further mesh refinement between the two finest grids, while the dominant mode remains unchanged; they are not presented as evidence of a formal convergence order.

Table 2. Mesh-refinement check for the reference case $D_v = 10$.

| N | h | $A_u(200)$ | dominant mode |
|---|---|---|---|
| 64 | 0.981748 | 0.865810 | 11 |
| 128 | 0.490874 | 0.870550 | 11 |
| 256 | 0.245437 | 0.870650 | 11 |

### *Time-integration sensitivity*

Although the BDF integrator is adaptive, sensitivity to temporal resolution can still be assessed by restricting its maximum internal step. For the reference case $D_v = 10$, $N = 128$, and the broadband perturbation, the computation was repeated with BDF maximum-step values $2, 1, 0.5$ and $0.25$ while retaining the same error tolerances. Table 3 shows that the final activator amplitude is stable to the displayed digits and that the dominant Neumann mode remains $n = 11$. Thus, within this tested range, the reported pattern selection is insensitive to the imposed BDF maximum-step bound.

Table 3. Temporal-resolution sensitivity for the reference case $D_v = 10$ ($N = 128$).

| maximum BDF step | $A_u(200)$ | dominant mode |
|---|---|---|
| 2.00 | 0.870550 | 11 |
| 1.00 | 0.870550 | 11 |
| 0.50 | 0.870550 | 11 |
| 0.25 | 0.870550 | 11 |

### *Sensitivity to the initial perturbation magnitude*

Linear Turing theory concerns infinitesimal perturbations, whereas a nonlinear computation must start from a finite disturbance. The controlled mode-11 perturbation used in the threshold experiment was therefore multiplied by factors $0.25, 0.5, 1$ and $2$. The purpose of this test is to assess sensitivity of the nonlinear saturation level and transient timing to the initial disturbance magnitude, not to establish mode selection, because the spatial mode is prescribed. Within the tested perturbation-amplitude range, the late-time amplitude and dominant mode are essentially unchanged.

Table 4. Sensitivity to the magnitude of the initial perturbation ($D_v = 10$, $N = 128$).

| factor | initial cosine coefficient in u | $A_u(200)$ | dominant mode |
|---|---|---|---|
| 0.25 | 0.0025 | 0.868640 | 11 |
| 0.50 | 0.0050 | 0.868642 | 11 |
| 1.00 | 0.0100 | 0.868642 | 11 |
| 2.00 | 0.0200 | 0.868642 | 11 |

### *Finite-domain controlled-mode experiments*

The continuous dispersion relation predicts a fastest-growing wavenumber, but a bounded Neumann domain admits only the discrete set $k_n = n\pi/L$. Domain size therefore acts as a spectral sampling parameter. To illustrate this effect in a controlled manner, simulations were repeated on $L = 10\pi$, $15\pi$, $20\pi$ and $25\pi$ while keeping the mesh width approximately constant and exciting one admissible mode close to the fastest-growing continuous wavenumber. Table 5 records the imposed admissible mode and the resulting final mode. This experiment assesses the persistence of the corresponding finite-domain pattern; it should not be interpreted as an independent competition test among all modes. The broadband reference experiment above provides that separate mode-selection check.

Table 5. Domain-size dependence in the controlled Neumann-mode experiments.

| L | N | imposed mode | final mode | $k_n$ | $A_u(200)$ |
|---|---|---|---|---|---|
| $10\pi$ | 64 | 6 | 6 | 0.6000 | 0.804252 |
| $15\pi$ | 96 | 8 | 8 | 0.5333 | 0.861681 |
| $20\pi$ | 128 | 11 | 11 | 0.5500 | 0.868642 |
| $25\pi$ | 160 | 14 | 14 | 0.5600 | 0.865209 |

### *Interpretation in the context of Turing-pattern literature*

The numerical evidence should be interpreted within the classical diffusion-driven instability framework rather than as a claim that the Schnakenberg kinetics themselves are new. Turing (1952:46) described how reaction and diffusion can amplify small spatial irregularities in an initially

nearly homogeneous system and thereby initiate pattern formation. Maini, Painter and Chau (1997:3601) reviewed spatial pattern formation in chemical and biological reaction–diffusion systems and discussed diffusion-driven instability as a mechanism for generating spatial structure. In the present chapter, the homogeneous equilibrium, Jacobian matrix, and instability conditions for the adopted Schnakenberg system are derived directly from the governing equations.

A further point emphasized by modern Turing-system studies is that the classical two-species mechanism is a starting point rather than a complete description of biological morphogenesis. Changes in geometry, growth, anisotropy, network architecture and kinetic structure can modify the accessible instability region and the patterns ultimately observed. For the present chapter, these broader developments motivate a deliberately controlled benchmark: geometry, boundary conditions and kinetics are fixed so that the correspondence between the dispersion relation and the computed dominant mode can be examined without additional modelling effects. This makes the experiment suitable as a controlled numerical study and also clarifies which conclusions are specific to the chosen one-dimensional setting. Woolley, Krause and Gaffney (2021:7) formulated the classical Turing instability conditions in terms of kinetic stability, diffusion, and the admissible spatial wavenumbers of the domain.

### *Spectral growth-rate comparison*

The agreement of the final dominant mode with the dispersion relation can be examined more closely by evaluating the larger eigenvalue of the linearized modal matrix at every admissible Neumann wavenumber in the unstable band. For the reference parameters, modes $n = 9$ through $n = 14$ satisfy the continuous instability condition. Table 6 lists their physical wavenumbers and linear growth rates. The largest value occurs at $n = 11$, with $n = 12$ only slightly smaller. Thus the numerically observed mode $n = 11$ is not simply one member of a broad unstable set; it is the admissible mode with the largest linear amplification rate on the chosen finite domain.

Table 6. Linear growth rates of the admissible unstable Neumann modes for $D_v = 10$.

| **mode n** | $k_n$ | $k_n^2$ | **largest modal growth rate** |
|---|---|---|---|
| 9 | 0.4500 | 0.2025 | 0.003060 |
| 10 | 0.5000 | 0.2500 | 0.041781 |
| 11 | 0.5500 | 0.3025 | 0.056484 |
| 12 | 0.6000 | 0.3600 | 0.053167 |
| 13 | 0.6500 | 0.4225 | 0.035315 |
| 14 | 0.7000 | 0.4900 | 0.005183 |

This calculation also clarifies why a finite-domain computation need not reproduce the continuous maximizer exactly. The dispersion curve is defined

for a continuum of wavenumbers, whereas the boundary-value problem samples that curve only at $k_n = n\pi/L$. Consequently, the relevant linear prediction for a numerical experiment is obtained in two stages: first identify the continuous unstable band and its maximizer, then evaluate the growth rates on the admissible eigenmodes of the chosen domain. Table 6 performs this discrete spectral comparison for $L = 20\pi$, while Table 5 illustrates how changing $L$ changes the available spectral sampling.

***Reproducibility, scope and limitations***

Several design choices were kept fixed deliberately. The reaction parameters, boundary conditions, deterministic perturbations and solver tolerances are stated explicitly, and the principal conclusions are checked against independent analytical quantities. In particular, the threshold $D_v, c$ is obtained from the Turing discriminant before simulation; the unstable interval is obtained from the roots of the modal determinant; and the reference wavelength-selection test uses a broadband perturbation rather than preselecting the theoretically fastest mode. The dominant spatial mode is then identified from a type-I discrete cosine transform (DCT-I), which is compatible with the endpoint-inclusive grid and the cosine eigenmodes associated with the homogeneous Neumann boundary conditions. These checks reduce the risk of interpreting a visually structured numerical solution as evidence of diffusion-driven instability without a corresponding spectral mechanism.

The study does not establish a new chemical reaction mechanism or a new general theorem for reaction–diffusion systems. It provides a controlled numerical investigation of the classical instability mechanism for a specific benchmark system. The Schnakenberg model has been studied analytically and numerically in the reaction–diffusion literature. Accordingly, all parameter classifications and numerical diagnostics reported here are recomputed for the specific equations, domain, boundary conditions, and discretization used in this chapter rather than presented as new properties of the model itself.

There are also limitations. First, one-dimensional geometry suppresses the competition among differently oriented wavevectors that is central to spots and stripes in two dimensions. Second, the homogeneous coefficients exclude spatial heterogeneity, anisotropy and domain growth, each of which can alter mode selection. Third, deterministic perturbations are useful for controlled numerical assessment but do not represent stochastic forcing. The computations cover selected parameter sets and a finite final time, and the mesh and BDF studies are sensitivity checks rather than a formal convergence analysis. Finally, linear theory predicts early-time growth rates and fastest-growing wavelengths but does not, by itself, determine the fully nonlinear saturated amplitude. The late-time profile is produced by nonlinear interactions after the perturbation has left the linear regime.

These limitations suggest natural extensions without changing the central comparison principle. A two-dimensional study could compare the annulus of unstable wavevectors predicted by the dispersion relation with a two-dimensional discrete spectrum; heterogeneous diffusion could be tested against a spatially varying linear operator; and random initial perturbations could be repeated across several seeds to distinguish robust spectral selection from realization-dependent transients. In each case, the key requirement would remain the same: the numerical pattern should be related quantitatively to an independently derived stability prediction rather than assessed only by visual appearance.

## RESULTS AND DISCUSSION

The refinement tests provide an additional numerical safeguard. Using the broadband perturbation, the dominant mode remains $n = 11$ as the spatial grid is refined from $N = 64$ to $N = 256$, and the final amplitudes on the two finest meshes are nearly identical. The BDF maximum-step sensitivity study likewise leaves the reported amplitude and dominant mode essentially unchanged. These checks indicate that the reference pattern-selection result is robust to the tested spatial resolutions and BDF maximum-step bounds; they are not presented as a formal convergence-order study.

The present computations are intentionally one-dimensional. This restriction makes the spectral prediction transparent because the Neumann modes are indexed by a single integer and can be compared directly with a discrete cosine spectrum. In two spatial dimensions, the same instability mechanism applies, but multiple wavevectors can have comparable growth rates and nonlinear interactions may generate spots, stripes, labyrinthine structures, or mixed patterns.

Overall, the results demonstrate the importance of interpreting numerical pattern formation together with the corresponding stability and finite-domain spectral analyses. The present framework therefore provides a consistent connection between Turing instability theory and the computed spatial patterns without attributing new properties to the Schnakenberg model itself. Future studies may extend this approach to heterogeneous coefficients, more complex reaction kinetics, and multidimensional domains.

## REFERENCES

Khan, F. M., Ali, A., Hamadneh, N. N., Abdullah and Alam, M. N. (2021). Numerical Investigation of Chemical Schnakenberg Mathematical Model. Journal of Nanomaterials, 2021, 9152972. https://doi.org/10.1155/2021/9152972

Maini, P. K., Painter, K. J. and Chau, H. N. P. (1997). Spatial pattern formation in chemical and biological systems. *Faraday Transactions*, *93*, 3601–3610. https://doi.org/10.1039/A702602A

Schnakenberg, J. (1979). Simple chemical reaction systems with limit cycle behaviour. *Journal of Theoretical Biology*, *81*(3), 389–400. https://doi.org/10.1016/0022-5193(79)90042-0

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *237*(641), 37–72. https://doi.org/10.1098/rstb.1952.0012

Wei, J. and Winter, M. (2008). Stationary multiple spots for reaction–diffusion systems. *Journal of Mathematical Biology*, *57*, 53–89.

Woolley, T. E., Krause, A. L. and Gaffney, E. A. (2021). Bespoke Turing systems. Bulletin of Mathematical Biology, 83, 41. https://doi.org/10.1007/s11538-021-00870-y

# Data Preprocessing, Descriptive Analysis and Measures of Association for Circular Data: A Comparative Examination of Linear and Circular Approaches

**Tuğba SÖKÜT AÇAR**[1]

1- Assoc. Prof. Dr.; Çanakkale Onsekiz Mart University, Faculty of Science, Department of Statistics, Çanakkale, Türkiye; t.sokut@comu.edu.tr; ORCID: 0000-0002-4444-1671.

## ABSTRACT

Measurements such as wind direction, the time of day and the day of the year are circular data defined on the unit circle. In such data, the starting and end points coincide, and the choice of origin is arbitrary. Classical statistical methods developed for linear variables may therefore yield misleading results when applied to circular data. This chapter addresses the preparatory stages that should precede the modelling of data sets containing a circular variable: data preprocessing, descriptive statistics, graphical display and the measurement of association between variables. For each stage, the differences between linear and circular approaches are explained theoretically and illustrated with publicly available meteorological data. The findings show that using the arithmetic mean to aggregate hourly data to the daily scale leads to systematic errors, particularly on days dominated by northerly winds. Moreover, the linear mean and the Pearson correlation changed substantially, with the correlation even reversing its sign, merely as a result of changing the origin. In contrast, the circular mean and the circular–linear correlation coefficient yielded consistent results that were independent of the origin. For bimodal distributions, the mean resultant length alone was found to be an inadequate summary, with the rose diagram and axial statistics providing complementary information. It is therefore recommended that, before a circular variable is entered into a regression model, the preprocessing, descriptive and association analyses be carried out using circular methods and that the results be interpreted together with graphical displays.

*Keywords – Circular data, Directional statistics, Wind direction, Circular–linear correlation, Data preprocessing*

## INTRODUCTION

The first step in any statistical analysis is to identify correctly the scale and space on which the data are defined. Linear variables are defined on the real line: their values can be ordered from smallest to largest, and the distance between two values is measured by the absolute value of their difference. Some measurements, however, are periodic in nature, in the sense that the scale returns to its starting point once a certain value is reached. Compass directions, the time of day, the day of the week and the day of the year are examples of such measurements. Data of this kind are represented as points on the circumference of the unit circle and are referred to as circular or directional data (Fisher, 1993; Mardia and Jupp, 2000).

The defining feature of circular data is that distance in the numerical representation does not coincide with actual distance. The values 350° and 10° differ by 340 units numerically, yet they are only 20° apart on the circle. The arithmetic mean of these two observations is 180°, which points in exactly the

opposite direction, namely south. Both observations, however, lie close to north, and the correct mean is 0° (Figure 1). Similarly, the location of the origin (0°) and the sense of rotation (clockwise or anticlockwise) are matters of convention. In meteorology, for instance, wind direction is measured clockwise from north, whereas in mathematics angles are measured anticlockwise from east. Linear statistics are sensitive to these arbitrary choices, whereas circular statistics are defined so that their results do not depend on them (Jammalamadaka and SenGupta, 2001).

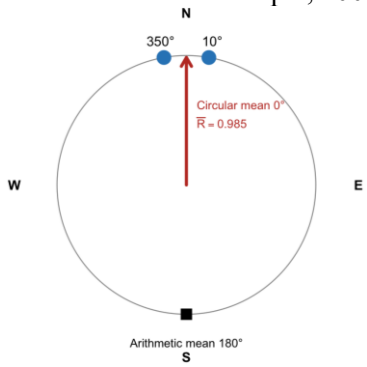



Figure 1: Arithmetic and circular means of two observations at 350° and 10°

Circular data arise in many scientific fields. Examples of circular measurements include the migration and orientation behaviour of animals in biology, the directions of wind and cloud movement in meteorology, the strike of rock strata and the dip directions of faults in the earth sciences, and the directions of waves and currents in marine science (Batschelet, 1981; Lark, Clifford, and Waters, 2014). Time may also exhibit a circular structure. For example, the arrival times of patients at an emergency department can be studied over a 24-hour cycle, and the seasonal incidence of diseases over a 365-day cycle (Pewsey, Neuhäuser, and Ruxton, 2013). Circular data are also frequently obtained in psychology and cognitive science, for instance in direction-estimation and timing tasks (Cremers and Klugkist, 2018).

The use of circular variables in regression models has become increasingly common in recent years. In such models, the response variable, the explanatory variables or both may be circular (Kim and SenGupta, 2015; SenGupta and Ugwuowo, 2006). In air-quality studies in particular, wind direction is an important explanatory variable that governs the transport of pollutants (Jammalamadaka and Lund, 2006). A circular explanatory variable is usually entered into the model through its cosine and sine components. When these components are correlated with the other explanatory variables, multicollinearity may arise. In such cases, biased estimators such as the ridge and Liu estimators can be used as alternatives to the ordinary least squares method (Söküt Açar, 2024).

The reliability of a model, however, does not depend solely on the estimation method. The preparatory stages carried out before the model is fitted also

shape the results. These stages include how the data are cleaned, how missing and erroneous records are handled, how the variables are described and how the associations among them are measured. When a circular variable is involved, each of these stages requires decisions that differ from those made for linear data. Although studies on circular regression models are growing rapidly in the literature, practice-oriented resources that address these preparatory stages systematically, setting linear and circular approaches side by side, remain limited. In the Turkish literature, studies have mostly focused on specific tests or applications (Peker and Bacanlı, 2009; Ser, 2014).

The aim of this chapter is to present, in a systematic manner, the steps that should be followed before modelling data sets containing a circular variable, with emphasis on the differences between linear and circular approaches. The chapter first explains the representation of circular data and then addresses, in turn, data preprocessing, descriptive statistics, graphical display and measures of association. These steps are then applied to a publicly available meteorological data set, and the chapter concludes with recommendations for the transition to modelling.

## REPRESENTATION OF CIRCULAR DATA

A circular observation is a point on the unit circle or, equivalently, a unit vector pointing to that point. Observations are usually recorded as angles, measured either in degrees (0°–360°) or in radians (0–2π). Before trigonometric functions can be applied, measurements in degrees are first converted to radians:

$$\theta_{\text{rad}} = \theta_{\text{deg}} \times \frac{\pi}{180} \qquad (1)$$

Each angle can also be expressed by its Cartesian coordinates on the unit circle:

$$\mathrm{u}_i = (\cos\theta_i,\ \sin\theta_i), \qquad i = 1, \dots, n \qquad (2)$$

This representation underlies almost all methods of circular statistics. When one works with the cosine and sine components rather than with the angle itself, the fact that 0° and 360° denote the same point no longer poses a problem, because the two angles have identical components. The same representation also determines how a circular explanatory variable is entered into circular–linear regression models (Mardia and Jupp, 2000).

The measurement convention should be stated explicitly when working with circular data. Meteorological wind direction indicates the direction from which the wind blows; it starts at north (0°) and increases clockwise. Accordingly, 90° denotes a wind blowing from the east and 180° a wind blowing from the south. The mathematical angle, by contrast, starts at east and increases anticlockwise. The two conventions are related by the transformation $\varphi = (90° - \theta) \bmod 360°$. Since circular statistics do not

depend on the choice of origin, the convention affects how the results are expressed but not their substance. Nevertheless, stating the convention used in reporting is necessary for the results to be read correctly.
For some directional measurements, a direction cannot be distinguished from its opposite. For example, 30° and 210° denote the same axis when describing the strike of a fault line or the alignment of a road. Such data are called axial data. For axial data, circular methods are applied to the doubled angles (2θ), and the resulting mean direction is then halved (Fisher, 1993). This transformation is also a useful tool for describing wind data concentrated in two opposite directions.

## DATA PREPROCESSING: DIFFERENCES BETWEEN LINEAR AND CIRCULAR APPROACHES

Data preprocessing involves identifying erroneous, missing and inconsistent records, bringing the variables into a form suitable for analysis and, where necessary, aggregating the data to a different time scale. Established practices for linear variables cannot be transferred directly to circular variables. The main issues specific to circular variables are discussed below and summarised in Table 1.

### *Coding and boundary values*

For a circular variable, 0° and 360° denote the same direction. The presence of both values in a data set is not an error, but for consistency all angles should be mapped to the interval [0°, 360°) using the operation $\theta \bmod 360°$. For a linear variable, values outside the admissible range (e.g., negative humidity) are flagged as errors. For a circular variable, by contrast, a value such as 370° usually means 10° and can be retained after transformation.

### *Calm winds and undefined direction*

When wind speed is zero, wind direction is undefined. In many measurement systems, calm conditions are coded as 0 in the direction variable. If these records are left unchanged, the frequency of northerly winds is artificially inflated and the mean direction is pulled towards north. An analogous problem for linear variables is the coding of missing values with numbers such as 0 or 999. But for circular data, such coding is much harder to detect, because 0° is a valid direction. Wind direction should therefore always be examined together with wind speed, calm records should be excluded from the directional analysis, and their number should be reported separately.

### *Missing values*

The mechanism of missingness (missing completely at random, missing at random, or missing not at random) is examined in the same way for circular and linear variables (Little and Rubin, 2019). The difference arises at the

imputation stage. For a linear variable, a missing value may be imputed using the mean, median, or linear interpolation between neighbouring observations. For a circular variable, a circular measure of location, such as the circular mean, may be used when a simple location-based imputation is appropriate. For interpolation, however, the cosine and sine components should be interpolated separately rather than the angular values themselves. For example, direct linear interpolation between 350° and 10° gives 180°, which is inappropriate because these directions are close to each other on the circle. In contrast, interpolation of the cosine and sine components gives a direction of 0°, which correctly reflects their circular proximity.

***Time stamps***

In hourly or sub-hourly data, daylight saving time causes some days to contain 23 hours and others 25 hours. If time stamps are read without time zone information, the repeated hour may appear as a duplicate record and the skipped hour as a missing one. These corrections become even more important when time itself is to be used as a circular variable (e.g., the time of day or the day of the year).

***Temporal aggregation***

When daily or monthly values are derived from hourly observations, the arithmetic mean is used for linear variables. For a circular variable, however, the arithmetic mean yields an incorrect result whenever the observations fall on both sides of the 0°/360° boundary. The correct approach is to compute the circular mean for each period. For wind data, a vector mean can also be computed by weighting the direction in each hour by the corresponding wind speed:

$$\bar{\theta}_w = \operatorname{atan2}\left(\sum_i w_i \sin\theta_i, \sum_i w_i \cos\theta_i\right) \qquad (3)$$

Here, $w_i$ denotes the weights (e.g., wind speeds) and $\operatorname{atan2}(\cdot,\cdot)$ is the two-argument arctangent function. It is also advisable to retain the mean resultant length of each period during aggregation. This quantity indicates how stable the direction was within the period and can serve as a weight or quality indicator in subsequent analyses.

***Outliers and transformations***

For linear variables, outliers are identified with tools such as the box plot and the interquartile range, and logarithmic or square-root transformations are often used to make a distribution more symmetric. For a circular variable, the notions of a smallest and a largest value are meaningless, so rules based on the box plot and the interquartile range cannot be applied. Circular outliers are identified instead through the circular distance of an observation from the mean direction or through criteria specific to circular regression (Abuzaid,

Hussin, and Mohamed, 2013). The transformations used for linear variables have no circular counterpart. The main admissible transformations are rotation (changing the origin) and, for axial data, doubling of the angles.

Table 1: Differences between linear and circular approaches in data preprocessing

| **Issue** | **Linear variable** | **Circular variable** |
|---|---|---|
| Range and boundary values | Out-of-range values are flagged as errors | Angles are mapped to [0°, 360°) via mod 360; 0° and 360° denote the same direction |
| Undefined values | Generally, not applicable | Direction is undefined when wind speed is zero; calm records are excluded |
| Imputation of missing values | Mean, median, linear interpolation | Circular mean; interpolation on the cos and sin components |
| Temporal aggregation | Arithmetic mean | Circular mean or speed-weighted vector mean; period-wise $\bar{R}$ is retained |
| Outliers | Box plot, interquartile range, z-scores | Circular distance from the mean direction; criteria specific to circular regression |
| Transformations | Logarithm, square root, standardisation | Rotation only; doubling of angles for axial data |

## DESCRIPTIVE STATISTICS

### *Descriptive Measures for Linear Variables*

The central tendency of a linear variable is described by the arithmetic mean and the median, and its dispersion by the standard deviation, the interquartile range and the range. The shape of the distribution is summarised by the coefficients of skewness and kurtosis, and conformity to the normal distribution is assessed with tests such as the Shapiro–Wilk test. Since these measures are well known, they are not discussed in detail here. It should be emphasised, however, that all of them rely on the ordered structure of the real line. The mean is the point that minimises the sum of squared distances, the median depends on the ordering of the observations, and the minimum and maximum correspond to the two ends of that ordering. Because circular data lack this structure, circular counterparts of these measures need to be defined separately.

***Descriptive Measures for Circular Variables***

*Mean direction and mean resultant length:* Let $\theta_1, \ldots, \theta_n$ denote circular observations. The means of the components of the corresponding unit vectors are computed as

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} \cos\theta_i, \qquad \bar{S} = \frac{1}{n}\sum_{i=1}^{n} \sin\theta_i \qquad (4)$$

Following Mardia and Jupp (2000), the mean resultant length and the mean direction are defined as

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}, \qquad \bar{\theta} = \operatorname{atan2}(\bar{S}, \bar{C}) \qquad (5)$$

The atan2 function selects the correct quadrant according to the signs of $\bar{C}$ and $\bar{S}$, so that no case-by-case distinction by quadrant is needed when computing the mean direction. The mean direction is undefined when $\bar{C} = \bar{S} = 0$. $\bar{R}$ takes values between 0 and 1. It equals 1 when all observations point in the same direction and approaches zero when the observations are spread evenly around the circle. A small value of $\bar{R}$ alone, however, does not imply that the distribution is uniform. In a bimodal distribution concentrated in two opposite directions, the vectors also cancel each other out, and $\bar{R}$ is small. $\bar{R}$ should therefore always be interpreted together with graphical displays.

*Circular variance and standard deviation:* Following Fisher (1993), the circular variance and the circular standard deviation are computed as

$$V = 1 - \bar{R}, \qquad v = \sqrt{-2\ln\bar{R}} \qquad (6)$$

$V$ takes values between 0 and 1. The circular standard deviation is expressed in radians but may be reported in degrees. Unlike in the linear case, the circular standard deviation is not the square root of the circular variance. As the data become more concentrated, $v$ approaches the ordinary standard deviation, whereas it increases without bound as the data spread around the circle.

*Circular median:* The circular median is the direction that minimises the sum of circular distances to the observations:

$$\tilde{\theta} = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} (\pi - |\pi - |\theta_i - \alpha||) \qquad (7)$$

Geometrically, the circular median is the endpoint, on the side where the observations are concentrated, of the diameter that divides the observations into two halves, with half of them on each side (Fisher, 1993). It is less sensitive to outliers than the mean direction.

*Circular skewness and kurtosis:* Symmetry and peakedness about the mean direction are measured by coefficients derived from the second trigonometric moments:

$$\bar{b}_2 = \frac{1}{n}\sum_{i=1}^{n} \sin 2(\theta_i - \bar{\theta}), \qquad \bar{a}_2 = \frac{1}{n}\sum_{i=1}^{n} \cos 2(\theta_i - \bar{\theta}) \qquad (8)$$

Here, $\bar{b}_2$ is the coefficient of skewness and $\bar{a}_2$ the coefficient of kurtosis (Pewsey et al., 2013). A value of $\bar{b}_2$ close to zero indicates symmetry about the mean direction, while a value of $\bar{a}_2$ close to zero indicates the absence of marked peakedness about the mean direction. For distributions concentrated in two opposite directions, $\bar{a}_2$ may be positive, because the transformation $2(\theta_i - \bar{\theta})$ maps opposite directions onto the same point.

*Rotation invariance:* The most important difference between circular and linear measures is that circular measures are unaffected by the choice of origin. If all observations are rotated by the same angle $\delta$, the mean direction rotates by exactly $\delta$, while $\bar{R}$, $V$, $\nu$, $\bar{a}_2$ and $\bar{b}_2$ remain unchanged. The linear mean and standard deviation do not have this property. When the origin is changed, some of the observations cross to the other side of the 0°/360° boundary, and the linear measures change arbitrarily.

***Circular Probability Distributions and Goodness-of-Fit Tests***

The probability distribution of a circular variable describes how probability mass is distributed around the unit circle. The principal continuous circular distributions are the circular uniform, cardioid, wrapped normal, wrapped Cauchy and von Mises distributions. In circular statistics, the role of the normal distribution is played by the von Mises distribution:

$$f(\theta \mid \mu, \kappa) = \frac{\exp\{\kappa\cos(\theta - \mu)\}}{2\pi I_0(\kappa)}, \qquad 0 \le \theta < 2\pi \qquad (9)$$

Here, $\mu$ is the mean direction, $\kappa \ge 0$ is the concentration parameter, and $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero. When $\kappa = 0$, the distribution reduces to the circular uniform distribution. As $\kappa$ increases, the probability mass becomes increasingly concentrated around $\mu$. For large values of $\kappa$, the distribution approaches a normal distribution with variance $1/\kappa$ (Mardia and Jupp, 2000). The maximum likelihood estimate of $\kappa$ is the solution of the equation $\mathrm{A}(\hat{\kappa}) = I_1(\hat{\kappa})/I_0(\hat{\kappa}) = \bar{R}$. In practice, the approximate formulae given by Fisher (1993) can be used.

*Rayleigh test:* The first hypothesis to be tested for circular data is whether the data are concentrated in any particular direction. The null hypothesis $H_0$ states that the distribution is uniform, and the alternative $H_1$ that the distribution is unimodal and concentrated in one direction. The test statistic is $Z = n\bar{R}^2$, and the approximation

$$p \approx \exp\left[\sqrt{1 + 4n + 4(n^2 - R_n^2)} - (1 + 2n)\right], \qquad R_n = n\bar{R} \qquad (10)$$

can be used to obtain the $p$ -value (Zar, 2010). Two issues are important when interpreting the Rayleigh test. First, rejection of $H_0$ does not imply that the data follow a von Mises distribution; it indicates only a departure from uniformity in the form of concentration. Second, the test is powerful against unimodal concentration. For bimodal distributions concentrated in opposite

directions, however, $\bar{R}$ is small, and the hypothesis of uniformity may therefore wrongly fail to be rejected.

*Goodness-of-fit tests:* Whether the data follow a particular distribution, such as the von Mises distribution, can be tested with Watson's $U^2$ test (Watson, 1961). This test is the circular counterpart of the Cramér–von Mises test, adapted so that its result does not depend on the choice of origin. Kuiper's test is used for a similar purpose (Pewsey et al., 2013). Critical values and software implementations of these tests are given in detail by Pewsey et al. (2013) and Jammalamadaka and SenGupta (2001).

## GRAPHICAL DISPLAY

The distribution of a linear variable is examined with the histogram, the box plot, the density plot and the violin plot, which combines the last two. When a circular variable is displayed with these graphs, the two ends of the scale (0° and 360°) are drawn far apart, although they are in fact adjacent. A distribution concentrated around north thus appears in a linear histogram as two separate piles at opposite ends, giving the false impression of bimodality. The main graphical displays for circular data are as follows (Pewsey et al., 2013):

- ***Raw circular data plot:*** Each observation is marked as a point on the circumference of the unit circle, and observations in the same direction are stacked outwards. This plot is suitable for small samples.
- ***Circular histogram:*** The circle is divided into angular bins of equal width, and the number of observations in each bin is represented by a bar extending outwards from the circle.
- ***Rose diagram:*** Each bin is drawn as a circular sector (wedge) radiating from the centre. Since the area of a wedge is proportional to the square of its radius, the radius should be drawn proportional to the square root of the frequency; otherwise, the visual weight of high-frequency directions is exaggerated. In meteorology, this graph is known as the wind rose and is usually coloured by wind-speed classes.
- ***Mean direction vector:*** An arrow drawn from the centre towards the mean direction, with length proportional to $\bar{R}$ . When superimposed on a rose diagram, it shows both the mean direction and the strength of concentration at a glance.
- ***Circular–linear plots:*** The mean or median of a linear variable within each directional sector can be displayed with a polar bar chart. This makes it possible to examine, for example, in which directions pollutant concentrations or wind speeds are high. Cylindrical scatter plots, which place the linear variable on the vertical axis and the angle around a circle, can also be used.

When drawing circular graphs, the origin and the sense of rotation should be chosen in accordance with the convention of the data. For meteorological data, north should be at the top and rotation should be clockwise. If the bin width is

too narrow, the graph becomes noisy; if it is too wide, the underlying structure is obscured. Bin widths between 10° and 30° are reasonable for most applications.

## MEASURES OF ASSOCIATION BETWEEN VARIABLES

The coefficient used to measure association depends on the types of the variables involved. Different coefficients have been defined for two linear variables, for one circular and one linear variable, and for two circular variables (Table 2).

### *Linear–Linear Association*

The linear association between two linear variables is measured by the Pearson correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (11)$$

The coefficient $r_{xy}$ takes values between −1 and +1; its sign indicates the direction of the relationship and its absolute value the strength. When the assumption of normality is not met, or when the relationship is monotonic but not linear, Spearman's rank correlation coefficient may be preferred. High correlations among the explanatory variables are the first sign of multicollinearity in a regression model.

When a circular variable is treated as linear and entered into the Pearson coefficient, two problems arise. First, because the relationship is periodic on the circle, it cannot be captured by a linear function; a northerly wind, for example, may be recorded as either 5° or 355°. Second, and more importantly, the value of the coefficient depends on the arbitrary choice of origin. For the same data, correlations of different magnitude, and even of opposite sign, can be obtained simply by changing where 0° is placed.

### *Circular–Linear Association*

For the association between a linear variable $x$ and a circular variable $\theta$, the coefficient proposed by Mardia (1976) and Johnson and Wehrly (1977) is obtained from the correlations of $x$ with $\cos\theta$ and $\sin\theta$:

$$r_{x\theta}^2 = \frac{r_{xc}^2 + r_{xs}^2 - 2\,r_{xc}\,r_{xs}\,r_{cs}}{1 - r_{cs}^2} \qquad (12)$$

where $r_{xc} = \mathrm{corr}(x, \cos\theta)$, $r_{xs} = \mathrm{corr}(x, \sin\theta)$ and $r_{cs} = \mathrm{corr}(\cos\theta, \sin\theta)$. The quantity $r_{x\theta}^2$ is equal to the coefficient of determination of the multiple linear regression of $x$ on $\cos\theta$ and $\sin\theta$. It therefore takes values between 0 and 1 and has no sign; indeed, it is not meaningful to describe the association between a circular and a linear variable as "positive" or "negative". Its square root, $R_{x\theta}$, may also be reported. The coefficient is invariant under rotation, because rotating $\theta$ replaces $\cos\theta$ and $\sin\theta$ by linear combinations of

themselves, which leaves the coefficient of determination unchanged. The hypothesis $H_0: \rho_{x\theta}^2 = 0$ is tested with the statistic

$$\mathrm{F} = \frac{(n-3)\, r_{x\theta}^2}{1 - r_{x\theta}^2} \qquad (13)$$

When $x$ is normally distributed, the $F$ statistic follows, under $H_0$, an $F$ distribution with 2 and $n-3$ degrees of freedom (Mardia and Jupp, 2000). When the normality assumption is in doubt, a permutation test can be used instead. This coefficient summarises the association between $x$ and $\theta$ without regard to the other variables, that is, at the marginal level. In a multivariable model, the contribution of the circular variable can be assessed only through the coefficients of the cosine and sine terms within the model.

***Circular–Circular Association***

For the association between two circular variables $\alpha$ and $\beta$, the coefficient proposed by Fisher and Lee (1983) is based on pairwise comparisons of the observations:

$$\rho_T = \frac{\sum_{1 \le i < j \le n} \sin\left(\alpha_i - \alpha_j\right) \sin\left(\beta_i - \beta_j\right)}{\sqrt{\sum_{1 \le i < j \le n} \sin^2\left(\alpha_i - \alpha_j\right) \; \sum_{1 \le i < j \le n} \sin^2\left(\beta_i - \beta_j\right)}} \qquad (14)$$

The coefficient $\rho_T$ takes values between −1 and +1. Positive values indicate that when one variable rotates clockwise, the other tends to rotate clockwise as well, whereas negative values indicate a tendency to rotate in opposite directions. Significance can be assessed with a permutation test in which the observations of one of the variables are randomly reordered. An alternative coefficient, given by Jammalamadaka and SenGupta (2001), is based on the deviations of the observations from their respective mean directions:

$$r_c = \frac{\sum_{i=1}^{n} \sin\left(\alpha_i - \bar{\alpha}\right) \sin\left(\beta_i - \bar{\beta}\right)}{\sqrt{\sum_{i=1}^{n} \sin^2\left(\alpha_i - \bar{\alpha}\right) \; \sum_{i=1}^{n} \sin^2\left(\beta_i - \bar{\beta}\right)}} \qquad (15)$$

A test statistic that is asymptotically standard normal has been derived for this coefficient. Both coefficients are independent of the origins of the two variables. Examples of circular–circular association include the relationship between wind direction and the day of the year (seasonality) and that between wind direction and the previous day's wind direction.

Table 2: Measures of association by variable type

| Variable types | Coefficient | Range | Significance test |
|---|---|---|---|
| Linear–linear | Pearson's $r$ (Spearman's $r_s$) | [−1, 1] | $t$ test |
| Circular–linear | Mardia / Johnson–Wehrly $r^2_x\theta$ | [0, 1] | $F(2, n-3)$ or permutation test |
| Circular–circular | Fisher–Lee $\rho_T$; Jammalamadaka–SenGupta $r_c$ | [−1, 1] | Permutation test; large-sample normal approximation |

# APPLICATION

***Data Set and Preprocessing***

The empirical illustration uses the publicly available "Weather in Szeged " data set, shared on the Kaggle platform (Budincsevity, 2016). The data set contains hourly meteorological observations for Szeged, Hungary, covering the period 2006–2016, and comprises a total of 96,453 records. The variables considered in this study are temperature, relative humidity, wind speed, wind direction, visibility and atmospheric pressure. All analyses were carried out in the R programming language (R Core Team, 2025) using the RStudio integrated development environment. Circular statistics were computed directly from the formulae given in above, without recourse to any additional packages. A significance level of 0.05 was adopted for all tests.

A preliminary inspection revealed that the monthly mean resultant length of wind direction rose to 0.92 and 0.97 in the last two months of 2016, whereas it did not exceed 0.54 in any other month of any year. Over the same period, the standard deviation of hourly wind speed fell from 4.5–9.6 km/h in the other months to 2.3 and 1.7 km/h. Wind direction remaining almost constant for two months is climatologically implausible and suggests that the data for this period may have been infilled or incorrectly recorded. In light of this consistency check, the analyses were based on 2015, the most recent complete year in which no such problem was observed. The preprocessing steps applied and their outcomes are summarised in Table 3.

Table 3: Preprocessing steps and findings (2015, hourly data)

| Step | Finding | Action taken |
|---|---|---|
| Exact duplicate rows (full data set) | 24 rows | Removed |
| Time stamps (with time zone information) | 2015: 8,760 hours; 363 days with 24 hours, one day with 23 hours and one day with 25 hours (daylight saving transitions) | Days assigned using local time |
| Pressure values of 0 | 116 hours (12 days partially affected) | Coded as missing; daily means computed from available hours |
| Calm wind (speed = 0) | 152 hours, all with direction coded as 0° | Excluded from directional analysis |
| Direction = 0° and speed > 0 | 99 hours | Retained as genuine northerly winds |
| Direction = 360° | None | — |
| Units | Speed in km/h, humidity as a 0–1 proportion, direction in degrees | Speed converted to m/s, humidity to percent, direction to radians |
| Temporal aggregation | Hourly data | Linear variables: arithmetic mean; direction: circular mean ($n$ = 365 days) |

Table 3 highlights two findings specific to circular data. First, of the 251 hours in which wind direction was recorded as 0°, 152 were in fact calm hours (zero wind speed). Had these records not been removed, the frequency of northerly winds would have been artificially inflated. The remaining 99 hours correspond to genuine northerly winds and were retained. Second, missing pressure values were coded as zero. For a linear variable such values are easy to detect, since a pressure of 0 hPa is physically impossible. If the same coding were used for wind direction, however, it could easily go unnoticed, because 0° is a valid direction.

When the hourly data were aggregated to daily values, the arithmetic mean was used for the linear variables and the circular mean of the non-calm hours for wind direction. For comparison, the daily arithmetic mean of wind direction was also computed. The median circular difference between the daily mean directions obtained by the two methods is 13.3°, and the mean difference is 40.8°. The difference exceeds 45° on 125 of the 365 days and 90° on 67 days. The largest difference, 179.7°, corresponds to an almost exactly opposite direction. As shown in Figure 2, large deviations are concentrated on days when the circular mean is close to 0° or 360°, that is, on days dominated by northerly winds. On these days, the hourly observations fall on both sides of the boundary, and the arithmetic mean is pulled towards

the south. By contrast, the median difference between the speed-weighted vector mean (Equation 3) and the unweighted circular mean is only 4.2°. The median of the daily mean resultant lengths is 0.786, indicating that on most days wind direction was relatively stable within the day.

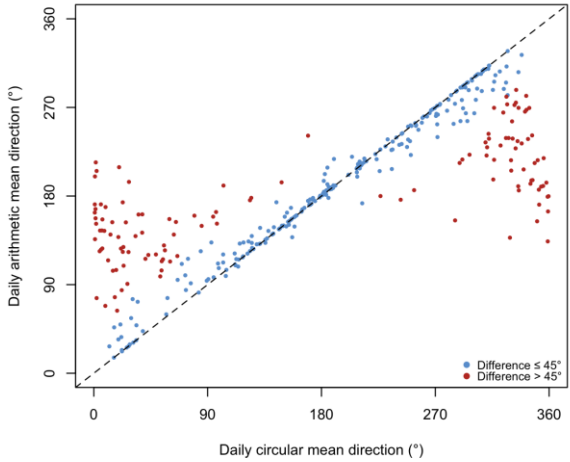



Figure 2: Comparison of the daily circular mean and the daily arithmetic mean of wind direction

**Descriptive Statistics for the Linear Variables**

Descriptive statistics for the linear variables aggregated to the daily scale are presented in Table 4, and their distributions are shown in Figure 3.

Table 4: Descriptive statistics for the linear variables (n = 365 days)

| Variable | Mean | SD | Median | Min – Max | Skewness | Kurtosis | SW (*p*) |
|---|---|---|---|---|---|---|---|
| Temperature (°C) | 12.31 | 8.58 | 11.85 | −7.18 – 29.22 | 0.135 | −1.056 | 0.965 ($p < .001$) |
| Relative humidity (%) | 73.23 | 14.06 | 73.08 | 41.62 – 98.50 | 0.023 | −1.057 | 0.969 ($p < .001$) |
| Wind speed (m/s) | 2.98 | 1.47 | 2.70 | 0.35 – 8.20 | 1.037 | 1.002 | 0.931 ($p < .001$) |
| Pressure (hPa) | 1,018.73 | 8.33 | 1,017.46 | 982.55 – 1,040.70 | 0.014 | 1.268 | 0.967 ($p < .001$) |
| Visibility (km) | 10.91 | 4.31 | 12.08 | 0.07 – 15.92 | −0.781 | −0.389 | 0.909 (p< .001) |

Note. Skewness and kurtosis are given as the third standardised moment and the excess kurtosis (0 for the normal distribution), respectively. SD = standard deviation; SW = Shapiro–Wilk test statistic.

The daily mean temperature ranges from −7.18 °C to 29.22 °C. The temperature distribution is symmetric, but its kurtosis coefficient is negative (−1.06). This flat shape results from the merging of two clusters corresponding to the winter and summer months (Figure 3). Relative humidity is likewise symmetric and flat. Wind speed is right-skewed (1.04), and several high values appear as outliers in Figure 3. These values were retained because they represent high-wind days rather than apparent measurement errors. Pressure

shows a few observations in the left tail, corresponding to low-pressure systems. Visibility, in contrast, piles up at its upper limit (approximately 16 km) and exhibits a left-skewed distribution. According to the Shapiro–Wilk test, none of the variables follows a normal distribution (all $p<0.001$). It should be borne in mind, however, that with a sample as large as $n = 365$, even small departures from normality tend to be statistically significant.

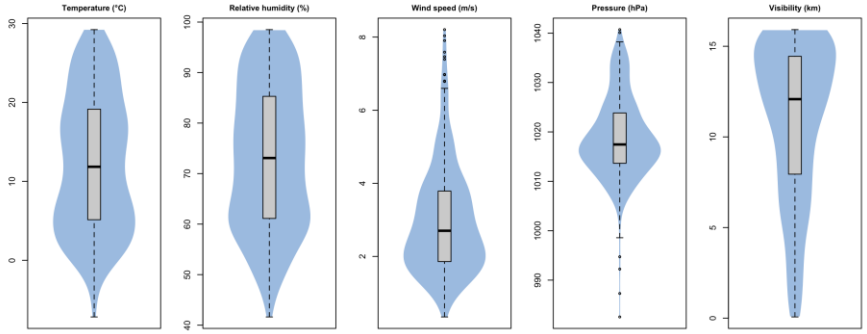



Figure 3: Violin and box plots of the linear variables

***Linear and Circular Descriptive Statistics for Wind Direction***

Table 5 compares the descriptive statistics for wind direction obtained by treating the variable first as linear and then as circular.

Table 5: Linear and circular descriptive statistics for wind direction

| **Measure** | **Daily (linear)** | **Daily (circular)** | **Hourly (linear)** | **Hourly (circular)** |
|---|---|---|---|---|
| Mean (direction) | 185.4° | 324.1° | 187.3° | 318.9° |
| Median | 183.9° | 312.5° | 182.0° | 303.0° |
| Standard deviation | 111.4° | 120.8° | 110.4° | 124.7° |
| Mean resultant length ($\bar{R}$) | — | 0.108 | — | 0.094 |
| Circular variance (V) | — | 0.892 | — | 0.906 |
| Skewness | −0.138 | $\bar{b}_2 = 0.092$ | — | $\bar{b}_2 = 0.150$ |
| Kurtosis | −1.260 | $\bar{a}_2 = 0.137$ | — | $\bar{a}_2 = 0.165$ |
| Concentration ($\hat{\kappa}$) | — | 0.218 | — | 0.188 |
| Rayleigh $Z$ ($p$) | — | 4.276 (0.014) | — | 75.32 (< 0.001) |
| Axial mean / $\bar{R}_2$ | — | 161.0° / 0.165 | — | 160.1° / 0.223 |

The linear mean (185.4°) implies that the wind blows, on average, from the south. As Figure 4 shows, however, south is not the prevailing direction. The

linear histogram (Figure 4a) displays two separate clusters, one near 0° and the other near 360°. The rose diagram (Figure 4b) reveals that these two clusters in fact constitute a single concentration in the north–northwest direction. The circular mean of the daily data is 324.1° (northwest), and the circular median is 312.5°. When the directions are classified into the eight principal compass points, the most frequently observed directions are north (64 days) and northwest (64 days), followed by southeast (48 days) and by south and west (45 days each). The least frequent direction is east (26 days).

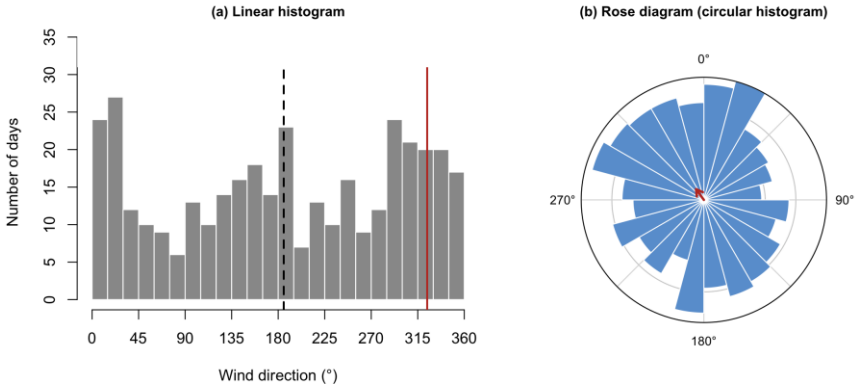



Figure 4: (a) Linear histogram and (b) rose diagram of daily wind direction. In panel (a), the dashed line indicates the linear mean (185.4°) and the red line the circular mean (324.1°). In panel (b), wedge radii are proportional to the square root of the frequencies, and the red arrow indicates the mean direction, with length proportional to the mean resultant length.

The mean resultant length is small ($\bar{R} = 0.108$), and the circular standard deviation takes a large value of 120.8°. The Rayleigh test rejects the hypothesis of uniformity at the 0.05 level ($Z = 4.276$; $p = 0.014$), but the concentration is weak ($\hat{\kappa} = 0.218$). Figure 4b shows that the small value of $\bar{R}$ reflects not only dispersion but also clustering in two opposite directions: the clusters in the northwest and southeast partly cancel each other out. Indeed, in the axial analysis based on doubled angles, the mean resultant length rises to $\bar{R}_2 = 0.165$, and the dominant axis is found to lie at approximately 161°–341° (south-southeast to north-northwest). This indicates that the circular mean alone is not an adequate summary and that the shape of the distribution should be assessed together with graphical displays and, where necessary, axial statistics. The positive kurtosis coefficient ($\bar{a}_2 = 0.137$) is also consistent with this bimodal structure.

Similar results are obtained when the hourly data are used (excluding calm hours, n = 8,608): the circular mean is 318.9°, $\bar{R} = 0.094$, and the axial mean resultant length is $\bar{R}_2 = 0.223$. For the hourly data, the p-value of the Rayleigh test is very small. However, because successive hourly observations are not independent, the assumption underlying the test is violated, and this result should be interpreted with caution.

To illustrate the fundamental difference between linear and circular measures, the measures were recomputed with the origin placed at 0°, 90°, 180° and 270° (Table 6). In each case, the resulting mean was converted back to the original convention (north = 0°) for comparability.

Table 6: Behaviour of the mean wind direction under changes of origin

| **Direction taken as 0°** | **Linear mean** | **Linear standard deviation** | **Circular mean** |
|---|---|---|---|
| 0° | 185.4° | 111.4° | 324.1° |
| 90° | 272.2° | 97.5° | 324.1° |
| 180° | 356.0° | 100.7° | 324.1° |
| 270° | 72.9° | 104.9° | 324.1° |

As Table 6 shows, merely changing the origin causes the linear mean to vary between 72.9° (east-northeast) and 356.0° (north). The linear standard deviation likewise fluctuates between 97.5° and 111.4°. The circular mean, in contrast, is 324.1° in every case. That an arbitrary choice such as the origin can determine the result for the same physical data is the clearest indication that linear measures are not valid for circular variables.

Monthly descriptive statistics are given in Table 7, and the monthly rose diagrams are shown in Figure 5.

Table 7: Monthly descriptive statistics for temperature, wind speed and wind direction

| **Month** | **Temperature (°C)** | **Wind speed (m/s)** | **Direction: linear mean ± SD** | **Direction: circular mean** | **$\bar{R}$** | **Circular SD** |
|---|---|---|---|---|---|---|
| January | 2.30 ± 3.68 | 3.63 ± 1.64 | 211.8° ± 96.9° | 232.4° | 0.155 | 110.6° |
| February | 2.42 ± 3.04 | 2.99 ± 1.51 | 168.6° ± 112.3° | 47.0° | 0.172 | 107.5° |
| March | 7.24 ± 2.64 | 3.98 ± 1.77 | 168.3° ± 113.5° | 358.8° | 0.141 | 113.4° |
| April | 11.61 ± 4.03 | 3.98 ± 1.88 | 212.5° ± 105.8° | 281.4° | 0.330 | 85.4° |
| May | 17.11 ± 2.68 | 3.17 ± 1.13 | 175.7° ± 118.8° | 348.4° | 0.153 | 111.0° |
| June | 20.50 ± 3.04 | 2.90 ± 0.97 | 232.7° ± 116.7° | 322.1° | 0.368 | 81.0° |
| July | 23.73 ± 3.72 | 2.53 ± 0.93 | 190.8° ± 116.2° | 325.4° | 0.238 | 97.1° |
| August | 23.78 ± 3.52 | 2.06 ± 0.73 | 148.9° ± 112.1° | 66.3° | 0.235 | 97.6° |

| Month | Temperature (°C) | Wind speed (m/s) | Direction: linear mean ± SD | Direction: circular mean | R̄ | Circular SD |
|---|---|---|---|---|---|---|
| September | 18.37 ± 4.60 | 3.19 ± 1.41 | 191.8° ± 117.5° | 340.0° | 0.200 | 102.8° |
| October | 10.62 ± 2.82 | 2.65 ± 1.49 | 147.9° ± 98.0° | 99.5° | 0.258 | 94.3° |
| November | 6.68 ± 4.46 | 2.50 ± 1.35 | 213.0° ± 110.1° | 287.7° | 0.386 | 79.0° |
| December | 2.66 ± 2.68 | 2.21 ± 0.93 | 164.2° ± 98.7° | 208.6° | 0.178 | 106.5° |

The monthly results show that the difference between the linear and circular means depends on the distribution of directions. The difference between the two means is smallest in January (approximately 21°), December (44°) and October (48°); in the remaining months it ranges from 69° to 173°. In March, May and September in particular, the prevailing direction is close to north (circular means of 358.8°, 348.4° and 340.0°, respectively). In these months, the linear means fall within the range 168°–192°, giving the entirely false impression that the wind blows from the south. In February, likewise, the linear mean is 168.6°, whereas the circular mean is 47.0°. The monthly mean resultant lengths lie between 0.14 and 0.39. Direction is most stable in November ($\bar{R} = 0.386$), June ($\bar{R} = 0.368$) and April ($\bar{R} = 0.330$). Northwesterly winds dominate in the summer months (June–July), whereas in August and October the mean direction shifts to the eastern half of the compass (Figure 5).

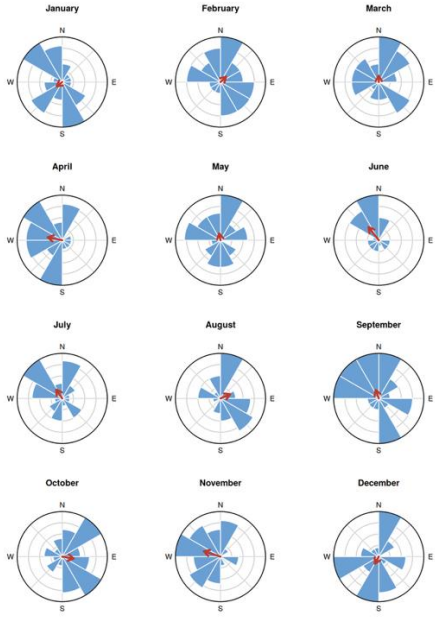



Figure 5: Monthly rose diagrams of daily wind direction (N = north, E = east, S = south, W = west; red arrow: mean direction vector).

### *Associations Between Variables*

Pearson correlation coefficients among the linear variables are presented in Table 8; for comparison, wind direction is also included, treated as a linear variable.

Table 8: Pearson correlation coefficients (n = 365; wind direction treated as a linear variable)

| | Temperature | Relative humidity | Wind speed | Pressure | Visibility | Wind direction |
|---|---|---|---|---|---|---|
| Temperature | 1 | | | | | |
| Relative humidity | −0.556** | 1 | | | | |
| Wind speed | −0.105* | −0.163** | 1 | | | |
| Pressure | −0.409** | 0.175** | −0.310** | 1 | | |
| Visibility | 0.603** | −0.859** | 0.178** | −0.263** | 1 | |
| Wind direction | −0.004 | −0.004 | 0.163** | −0.135* | 0.032 | 1 |

Note. *p < 0.05. **p <0.01.

There are strong associations among the linear variables. Significant correlations are observed between relative humidity and visibility ($r = -0.859$), temperature and visibility ($r = 0.603$), temperature and relative humidity ($r = -0.556$), and temperature and pressure ($r = -0.409$). If these variables were used together as explanatory variables in a regression model,

multicollinearity should be expected. When wind direction is treated as linear, only weak but significant associations with wind speed ($r = 0.163$) and pressure ($r = -0.135$) are found. The meaning of these coefficients, however, is unclear. The positive correlation with wind speed, for example, would require an interpretation with no physical counterpart, namely that "wind speed increases as the direction angle increases".
The association between wind direction and the linear variables was recalculated using Mardia's circular–linear correlation coefficient (Equation 12), and the results are presented in Table 9.

Table 9: Circular–linear correlations between wind direction and the linear variables

| **Variable** | $r_{xc}$ | $r_{xs}$ | $R$ | $R^2$ | $F$ | $p$ |
|---|---|---|---|---|---|---|
| Temperature | −0.022 | 0.049 | 0.053 | 0.0028 | 1.003 | 0.368 |
| Relative humidity | 0.025 | 0.020 | 0.034 | 0.0011 | 0.412 | 0.662 |
| Wind speed | 0.038 | −0.194 | 0.195 | 0.0381 | 14.339 | <0.001 |
| Pressure | 0.065 | 0.100 | 0.125 | 0.0155 | 5.709 | 0.004 |
| Visibility | 0.014 | −0.039 | 0.041 | 0.0017 | 0.599 | 0.550 |

Note. $F_{\text{critical}}(2,\ 362) = 3.021$

According to the circular–linear coefficients, wind direction is significantly associated with wind speed (R=0.195, p<0.001) and pressure (R=0.125, p=0.004), whereas no significant association is found with temperature, relative humidity or visibility. In this application, the linear and circular approaches led to similar conclusions as to which associations are significant. But this agreement is coincidental and depends on the choice of origin. To demonstrate this, the origin was varied from 0° to 355° in steps of 5°, and the Pearson correlation was recalculated in each case (Figure 6).

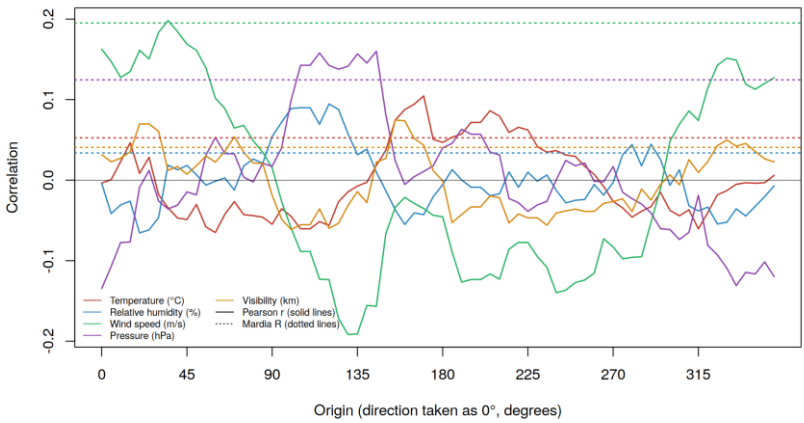



Figure 6: Pearson correlations (solid lines) and Mardia's circular–linear correlations (dotted lines) between wind direction and the linear variables as a function of the origin

Figure 6 clearly demonstrates how sensitive the linear correlation is to the arbitrary choice of origin. Depending on the origin, the Pearson correlation between wind speed and wind direction ranges from −0.192 to 0.198, changing sign in the process. The corresponding ranges are −0.135 to 0.160 for pressure and −0.065 to 0.105 for temperature. In other words, using the same data, one researcher might report a significant positive association between wind speed and direction, another a significant negative association, and a third no significant association at all. Mardia's coefficient, by contrast, does not depend on the origin (dotted horizontal lines). For some origins, the absolute value of the Pearson coefficient approaches or even exceeds Mardia's coefficient, whereas for others it falls to zero. The linear coefficient therefore provides consistent information neither about the direction nor about the strength of the association.

The structure of the association between wind direction and wind speed is examined in Figure 7. No clear pattern is visible on the linear axis (Figure 7a). The mean wind speeds for the eight principal compass directions (Figure 7b), however, show that northwesterly (3.40 m/s), westerly (3.36 m/s) and southeasterly (3.23 m/s) winds are stronger, whereas easterly winds are the weakest (1.99 m/s). The fact that the stronger winds are concentrated in two opposite sectors explains why the association cannot be summarised by a single linear slope. This structure suggests that, in a regression model, the circular variable may require second-harmonic terms ($\cos 2\theta$, $\sin 2\theta$) in addition to the first-harmonic terms ($\cos\theta$, $\sin\theta$).

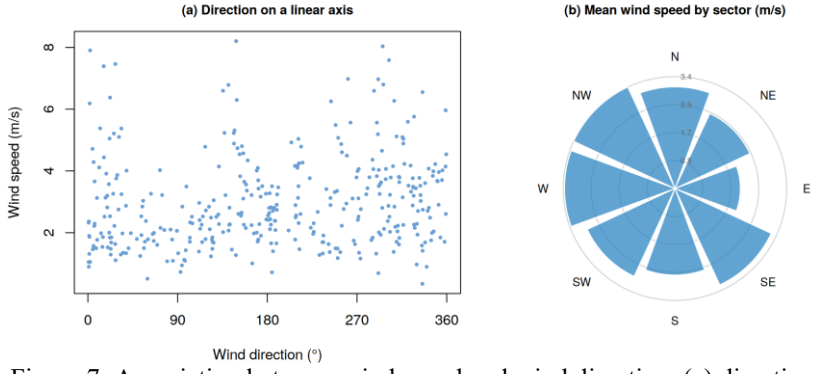



Figure 7: Association between wind speed and wind direction: (a) direction on a linear axis; (b) mean wind speed by the eight principal compass directions (N = north, NE = northeast, E = east, SE = southeast, S = south, SW = southwest, W = west, NW = northwest)

Finally, two questions concerning circular–circular association were examined. The first was whether wind direction follows a seasonal pattern. For this purpose, the day of the year was converted to an angle using the transformation $2\pi(g-1)/365$, where $g$ denotes the day of the year. The Fisher–Lee coefficient between wind direction and the day of the year is $\rho_T =$

0.006 (permutation test with 5,000 replications, $p = 0.092$), and the Jammalamadaka–SenGupta coefficient is $r_c = 0.021$ ($p = 0.691$). There is thus no consistent, one-directional seasonal rotation over the year. This result is also consistent with the irregular changes in the monthly mean directions over the year shown in Table 7. The second question was whether the wind direction on a given day is associated with that of the previous day. This association was found to be significant: $\rho_T = 0.120$ (permutation $p < 0.001$) and $r_c = 0.335$ ($p < 0.001$). Wind direction therefore exhibits day-to-day persistence. This finding also indicates that the daily observations are not independent and that the results of methods assuming independence, such as the Rayleigh test, should be interpreted with caution.

## DISCUSSION AND RECOMMENDATIONS FOR THE TRANSITION TO MODELLING

This chapter has examined the preparatory stages that precede the modelling of data sets containing a circular variable, comparing linear and circular approaches. The results of the application show that treating a circular variable as linear leads to different kinds of errors at every stage.

At the preprocessing stage, calm-wind records and genuine northerly winds, both coded with the same value (0°), could be distinguished only when examined together with wind speed. At the aggregation stage, use of the arithmetic mean produced errors exceeding 45° on more than one third of the days. These errors are not random: they arise systematically on days when the prevailing direction is north and make these days appear to be days with southerly winds. Once such an error enters a data set, it is carried over into all subsequent analyses and becomes difficult to detect.

At the descriptive stage, the linear mean and standard deviation varied arbitrarily with the choice of origin, whereas the circular measures remained invariant under rotation. It was also seen, however, that neither the circular mean nor the mean resultant length is sufficient on its own. In bimodal distributions, a small value of $\bar{R}$ does not mean that there is "no preferred direction"; in such cases, the rose diagram and axial statistics provide complementary information. At the stage of measuring association, the sign and significance of the linear correlation changed with the origin, whereas the circular–linear coefficient yielded consistent results.

In light of these findings, the following steps are recommended before fitting a regression model that includes a circular explanatory variable:

1. Specify the measurement convention (origin, sense of rotation and unit), and map all angles to the interval [0°, 360°).
2. Examine the circular variable together with the linear variable that gives it meaning (e.g., wind direction together with wind speed). Exclude undefined cases (calm winds) and report their number.

3. Identify missing values together with the way they are coded. If imputation is required, carry it out on the cosine and sine components of the circular variable.
4. When aggregating to a coarser time scale, use the circular or vector mean for the circular variable, and retain the mean resultant length of each period.
5. Describe the circular variable using the circular mean, median, $\bar{R}$, circular standard deviation, skewness and kurtosis. Test for concentration with the Rayleigh test and, where necessary, for goodness of fit with Watson's $U^2$ test.
6. Use a rose diagram or a circular histogram instead of a linear histogram. If multimodality is observed, compute axial statistics as well.
7. Measure the association between the circular variable and the linear variables with Mardia's coefficient, and the association between circular variables with the Fisher–Lee or Jammalamadaka–SenGupta coefficients.
8. Assess the linear and circular–linear correlations among the explanatory variables with regard to multicollinearity.
9. Make a preliminary assessment of the adequacy of the first-harmonic representation using direction-conditional plots (Figure 7b).

Following these steps, the circular variable can be entered into the regression model through its cosine and sine components. As Tables 8 and 9 show, however, strong associations among the linear explanatory variables, together with the associations between the trigonometric components and these variables, may render the design matrix ill-conditioned. In that case, the variance of the ordinary least squares estimator increases and the coefficient estimates become unstable. Multicollinearity diagnostics, such as the condition number and the eigenvalues, should be computed from the design matrix obtained after the trigonometric transformation. When multicollinearity is detected, biased estimators such as the ridge and Liu estimators can provide more stable estimates at the cost of a small bias (Söküt Açar, 2023; Hoerl and Kennard, 1970; Liu, 1993). The performance of these estimators in the circular–linear regression model was examined separately within the scope of the project on which this chapter is based (Söküt Açar, 2024).

This study has some limitations. The application was based on data from a single station for a single year, so the results cannot be generalised directly to other climatic regions. Only univariate and bivariate associations were examined in this chapter; the contribution of the circular variable to the response variable after controlling for the other variables can be assessed only with a multivariable model. Furthermore, the unusual patterns detected in some periods of the data set point to the need for thorough consistency checks before publicly available data sets are used.

## CONCLUSION

Although their numerical representation appears linear, circular data are defined on the circle and therefore require dedicated methods. As demonstrated in this chapter, applying linear methods to circular data produces misleading results not only at the modelling stage but also at the stages of data preprocessing, description, visualisation and the measurement of association. These errors are often silent: the computations run without error, and the results appear numerically plausible. The presence of a circular variable should therefore be taken into account from the very beginning of the analysis. The cosine–sine representation, the circular mean and the mean resultant length, the rose diagram and the circular–linear correlation are the essential tools in this process. A data set prepared with these tools provides a sound basis for circular–linear regression modelling and for the application of biased estimators under multicollinearity.

## ACKNOWLEDGEMENTS

This study was supported by the Scientific Research Projects Coordination Unit of Çanakkale Onsekiz Mart University under project number FHD-2023-4581.

## REFERENCE

Abuzaid, A. H., Hussin, A. G., and Mohamed, I. B. (2013). Detection of outliers in simple circular regression models using the mean circular error statistic. Journal of Statistical Computation and Simulation, 83(2), 269–277.

Batschelet, E. (1981). Circular statistics in biology. London, UK: Academic Press.

Budincsevity, N. (2016). Weather in Szeged 2006–2016 [Data set]. https://www.kaggle.com/datasets/budincsevity/szeged-weather

Cremers, J., and Klugkist, I. (2018). One direction? A tutorial for circular data analysis using R with examples in cognitive psychology. Frontiers in Psychology, 9, 2040. doi:10.3389/fpsyg.2018.02040

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

Fisher, N. I., and Lee, A. J. (1983). A correlation coefficient for circular data. Biometrika, 70(2), 327–332.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

Jammalamadaka, S. R., and Lund, U. J. (2006). The effect of wind direction on ozone levels: A case study. Environmental and Ecological Statistics, 13(3), 287–298.

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

Johnson, R. A., and Wehrly, T. E. (1977). Measures and models for angular correlation and angular–linear correlation. Journal of the Royal Statistical Society: Series B (Methodological), 39(2), 222–229.

Kim, S., and SenGupta, A. (2015). Inverse circular–linear/linear–circular regression. Communications in Statistics – Theory and Methods, 44(22), 4772–4782.
Lark, R. M., Clifford, D., and Waters, C. N. (2014). Modelling complex geological circular data with the projected normal distribution and mixtures of von Mises distributions. Solid Earth, 5(2), 631–639.
Little, R. J. A., and Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Hoboken, NJ: John Wiley and Sons.
Liu, K. (1993). A new class of biased estimate in linear regression. Communications in Statistics – Theory and Methods, 22(2), 393–402.
Mardia, K. V. (1976). Linear-circular correlation coefficients and rhythmometry. Biometrika, 63(2), 403–405.
Mardia, K. V., and Jupp, P. E. (2000). Directional statistics. Chichester, UK: John Wiley and Sons.
Peker, K. Ö., and Bacanlı, S. (2009). Yoğunlaşma parametresi bilinmediğinde dairesel verilerde ortalama yön için ardışık testin uygulanması [Application of the sequential test for the mean direction of circular data when the concentration parameter is unknown]. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 2(1), 9–18.
Pewsey, A., Neuhäuser, M., and Ruxton, G. D. (2013). Circular statistics in R. Oxford, UK: Oxford University Press.
R Core Team. (2025). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
SenGupta, A., and Ugwuowo, F. I. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data. Environmental and Ecological Statistics, 13(3), 299–309.
Ser, G. (2014). Directional data analysis and an application. Yüzüncü Yıl University Journal of Agricultural Sciences, 24(2), 121–126.
Söküt Açar, T. (2023). A new approach in modelling the circular data: Circular ridge estimator. Journal of Statistical Computation and Simulation, 93(4), 671–683.
Söküt Açar, T. (2024). Dairesel regresyon modellemesinde yeni bir yaklaşım: Liu tahmin edici [A new approach in circular regression modelling: The Liu estimator] (Project Report No. FHD-2023-4581). Çanakkale, Türkiye: Çanakkale Onsekiz Mart University, Scientific Research Projects Coordination Unit.
Watson, G. S. (1961). Goodness-of-fit tests on a circle. Biometrika, 48(1/2), 109–114.
Zar, J. H. (2010). Biostatistical analysis (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

# The Predicted *in Silico* Toxicological Profile of Malachite Green

**Nazlı TÜRKTEN*[1]**
**Yunus KARATAŞ[1]**

[1]Department of Chemistry/Faculty of Arts and Sciences, Kirsehir Ahi Evran University, Türkiye
*(nazli.turkten@ahievran.edu.tr) Email of the corresponding author

## ABSTRACT

Cationic triarylmethane dyes are aromatic xenobiotic compounds that contribute to environmental pollution and can cause serious health problems. Among these synthetic dyes, malachite green (MG) is commonly used in aquaculture as a parasiticide, as well as in food, textiles, and other industries for various applications. Many countries have strict regulations or outright bans on the use of MG due to its significant toxicity. Despite its effectiveness and low cost, the illegal use of this antibiotic, which also has antifungal properties in aquaculture, poses serious threats to both the ecological environment and human health. In this study, the predicted *in silico* toxicity risk assessment of MG was evaluated by StopTox, AdmetSAR 3.0, ADMETlab 3.0, and ProTox 3.0 applications. Qualitative predictions conducted with StopTox and AdmetSAR 3.0 classified MG as toxic, while AdmetSAR 3.0 indicated it was moderately toxic. The predicted median lethal dose ($LD_{50}$) value for MG was calculated as 80 mg/kg. ProTox 3.0 classified MG as a carcinogen, immunotoxin, and ecotoxicant, highlighting its implications for the blood-brain barrier (BBB).

*Keywords – ADMET 3.0, AdmetSAR 3.0, in silico, Malachite Green, ProTox 3.0, STopTox, Toxicity.*

## INTRODUCTION

Environmental pollution continues to be a significant risk for water ecosystems, human health, and various living organisms. Thus, thousands of individuals worldwide have become ill after consuming food products contaminated with harmful chemicals, making food safety a continuous concern due to its direct impact on public health. This pollution issue is primarily driven by population growth, industrial activity, and urbanization. A major contributor to this pollution is the improper disposal of toxic dyes, especially from wastewater generated by the textile, pharmaceutical, and dye manufacturing industries [1-5].

Malachite green (MG), a cationic triarylmethane dye, has been widely used across sectors for various purposes despite government regulations and legal prohibitions in most countries. In the textile industry, MG functions as a water-soluble basic dye with vibrant green and high color fastness, whereas in food production, it acts as a food-grade additive to prevent microbial growth. This aromatic xenobiotic compound is also used as a controversial agent in aquaculture, such as a parasiticide, disinfectant, and fungicide. This synthetic dye has been used in surgical materials as a disinfectant and prophylactic agent to prevent infections in surgical incisions due to its oxidative properties [1, 3-11]. It has been reported that MG is detected in food dyes and the flesh of fish that are raised in captivity for

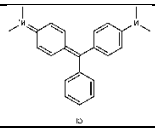
human consumption [6]. Due to disruptions in the food chain within aquatic ecosystems, the use of MG in aquatic food animals is heavily restricted or banned in many countries. However, its illegal use, even in diluted form in fish farms as an antimicrobial agent for fungal and bacterial diseases continues to rise because it is inexpensive, readily available, and effective. The presence of MG in aquaculture poses significant health risks by adversely affecting various human organs due to its teratogenic, genotoxic, carcinogenic, and immunosuppressive effects on consumers. The absorption of MG by fish and other animal tissues results in its reduced and colorless form, leucomalachite green, which can then enter the food chain and accumulate in various tissues [12-14].

Limited information is available on the toxicity profile of MG, although MG is known for its harmful effects. This *in silico* study focused on the predicted toxicological assessment of the MG dye using both qualitative (STopTox, AdmetSAR 3.0, and ADMETlab 3.0) and quantitative (ProTox 3.0) approaches. To our knowledge, no previous research has comparatively evaluated the predicted toxicological profile of MG. The acute toxicity, toxicity endpoints, and organ-specific toxicity profiles of this illicit food coloring agent were also investigated.

## MATERIALS AND METHOD

The predicted toxicological profile of MG was evaluated using StopTox, AdmetSAR 3.0, ADMETlab 3.0, and ProTox 3.0 . The STopTox (Systemic and Topical Chemical Toxicity) tool computes results from various human acute toxicity assays, collectively referred to as the "6-pack"[15]. ADMETlab 3.0 is a comprehensive platform for predicting the ADMET 3.0 (Absorption, Distribution, Metabolism, Excretion, and Toxicity) features of chemical compounds [16, 17]. AdmetSAR 3.0 is a highly advanced software tool for predicting the ADMET outcomes of chemical compounds [18-20]. ProTox 3.0, an open-source platform (https://tox.charite.de/protox3/), calculates toxicity and associated endpoints by using machine learning algorithms [21, 22]. The summary of chemical structure and properties of MG dye is presented in Table 1.

Table 1. The chemical structure and properties of MG dye

| Parameter | MG |
|---|---|
| IUPAC Name | [4-[[4-(dimethylamino)phenyl]-phenylmethylidene]cyclohexa-2,5-dien-1-ylidene]-dimethylazanium chloride |
| SMILES | CN(C)C1=CC=C(C=C1)C(=C2C=CC(=[N+](C)C)C=C2)C3=CC=CC=C3.[Cl-] |
| Molecular formula | $C_{23}H_{25}ClN_2$ |
| Chemical structure |  |
| Molar mass, g/mol | 364.9 |

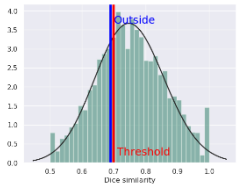
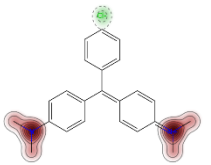
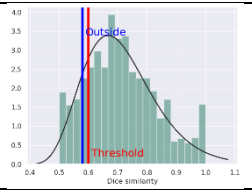
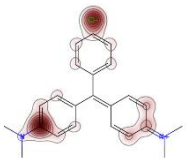
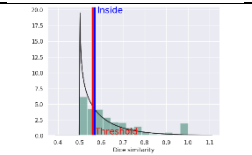
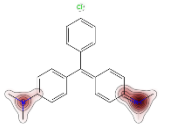
## RESULTS

The acute toxicity of MG was qualitatively predicted using ADMETlab 3.0, AdmetSAR 3.0 (Table 2), and STopTox (Table 3) to identify toxicophores and evaluate the overall toxicity profile. For the oral route, STopTox and ADMETlab 3.0 predicted high toxicity, whereas AdmetSAR predicted a moderate toxicity classification with a 62% probability. STopTox and ADMETlab 3.0 classified MG dye as toxic for acute inhalation. In contrast, AdmetSAR 3.0 predicted low toxicity (-)/slightly toxic with a 26% probability, suggesting lower confidence in this prediction and the necessity for further verification using additional tools. For dermal toxicity, STopTox and ADMETlab 3.0 predicted MG as toxic.

Table 2. Prediction of acute toxicity for MG dye using ADMETlab 3.0 and AdmetSAR 3.0

| | ADMETlab 3.0 | | AdmetSAR 3.0 | |
|---|---|---|---|---|
| **Type of acute toxicity** | **Prediction** | **Confidence (%)** | **Prediction** | **Confidence (%)** |
| Oral | High | 92.1 | Moderately Toxic (+) | 62.1 |
| Dermal | *NA* | *Not applicable* | Moderately Toxic (+) | 67.2 |
| Inhalation | High | 99.0 | Low Toxicity (-)/ Slightly Toxic | 26.0 |

Table 3. Prediction of acute toxicity for MG dye using STopTox

| Type of acute toxicity | Prediction | Confidence (%) | Applicability domain | Predicted toxicophore(s) |
|---|---|---|---|---|
| Oral | Toxic (+) | 81.0 |   |  |
| Dermal | Toxic (+) | 73.0 |   |  |
| Inhalation | Toxic (+) | 84.0 |   |  |

To quantitatively evaluate the acute toxicity of MG dye, ProTox 3.0 in silico tool was utilized. The predicted $LD_{50}$ value for MG was 80 mg $kg^{-1}$, classifying it as toxic when swallowed (class III). This value is consistent with the experimental finding of oral rat $LD_{50}$ 50 mg $kg^{-1}$ [23]. However, different experimental values of $LD_{50}$ for MG in rats ranged from 275 to 520 mg $kg^{-1}$ body weight [11, 24, 25]. The toxicological profile of MG, including its toxicity endpoints and organ toxicity features, is

summarized in Table 4. According to ProTox 3.0, MG is classified as a carcinogen, immunotoxin, and ecotoxicant, and it also has implications for the BBB.

Table 4. The predicted toxicity endpoints and organ toxicity profile of MG dye

| | **Toxicity endpoints** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Immunotoxicity** | **Mutagenicity** | **Cytotoxicity** | **BBB** | **Ecotoxicity** | **Clinical toxicity** | **Nutritional toxicity** | **Carcino-genicity** |
| **Prediction** | Active | Inactive | Inactive | Active | Active | Inactive | Inactive | Active |
| **Probability (%)** | 87 | 71 | 74 | 76 | 78 | 65 | 74 | 90 |
| | **Organ Toxicity** | | | | | | | |
| | **Hepatotoxicity** | **Neurotoxicity** | **Nephrotoxicity** | **Cardio-toxicity** | **Respiratory toxicity** | | | |
| **Prediction** | Inactive | Active | Inactive | Inactive | Active | | | |
| **Probability (%)** | 66 | 64 | 88 | 70 | 82 | | | |

## CONCLUSION

This study investigated, to our knowledge, for the first time the toxicity profile of MG dye, focusing on its acute toxicity, as well as endpoints and organ toxicity using various *in silico* methods. Computational predictions were conducted using various in silico tools, including qualitative (StopTox, AdmetSAR 3.0, ADMETlab 3.0) and quantitative (ProTox 3.0) methods. Quantitative analysis determined that the estimated $LD_{50}$ value for MG was 80 mg $kg^{-1}$. The differences in the predicted toxicity profiles among the *in silico* tools highlight the variability of computational predictions. Further research is necessary to assess the toxicity of MG, addressing the limitations of *in silico* methodologies. This could enhance predictive accuracy and expand the range of toxic effects evaluated.

## REFERENCES

[1] G.S.A. Varshan, S.K.R. Namasivayam, K.S. Sivasuriyan, S. R., Green Chitosan Nanoparticles for Malachite Green Dye Removal and Safety Evaluation, Chemical Engineering & Technology, 49 (2026) e70292.
[2] C. Ma, H. Kang, Q. Liu, R. Zhu, Z. Cao, Insight into potential toxicity mechanisms of melamine: An in silico study, Toxicology, 283 (2011) 96–100.
[3] M. Ahyad, H. Hardhienata, E.H. Hasdeo, S.A. Wella, F. Handayasari, H. Alatas, M.D. Birowosuto, A Novel Sensing Method to Detect Malachite Green Contaminant on Silicon Substrate Using Nonlinear Optics, Micromachines, 15 (2024) 1227.
[4] H.N. Ahamed, I. Y, I. Navabshan, M.Z. S, S. Ts, I. Jaleel, T.A. Lh, Investigating the toxicity of malachite green and copper sulfate in brine shrimp: In-vivo and computational study, Toxicology Reports, 13 (2024) 101811.
[5] A. Pandey, S. Kumar, N. Bithel, S. Kumar, M.A. Mir, Sustainable biodegradation of malachite green dye by novel non-pathogenic Pseudomonas aeruginosa ED24, World J. Microbiol. Biotechnol., 41 (2025) 44.
[6] A.O. de Almada Vilhena, K.M.M. Lima, L.F.C. de Azevedo, J.D. Rissino, A.C.P. de Souza, C.Y. Nagamachi, J.C. Pieczarka, The synthetic dye malachite green found in food induces cytotoxicity and genotoxicity in four different mammalian cell lines from distinct tissuesw, Toxicology Research, 12 (2023) 693–701.
[7] K. Chen, L. Zhu, Z. Du, X. Lan, K. Huang, W. Zhang, W. Xu, Docking-aided rational tailoring of a fluorescence- and affinity-enhancing aptamer for a label-free ratiometric malachite green point-of-care aptasensor, J. Hazard. Mater., 447 (2023) 130798.
[8] F. Joseph, Y.K. Agrawal, D. Rawtani, Behavior of malachite green with different adsorption matrices, Frontiers in Life Science, 7 (2013) 99–111.
[9] M. Kooravand, S. Asadpour, H. Haddadi, S. Farhadian, An insight into the interaction between malachite green oxalate with human serum albumin: Molecular dynamic simulation and spectroscopic approaches, J. Hazard. Mater., 407 (2021) 124878.

[10] R. Ma, W. Fang, H. Zhang, J. Sun, H. Su, T. Chen, K. Hu, Transcriptome analysis of zebra fish (Danio rerio) eggs following treatment with malachite green, Aquaculture, 514 (2020) 734500.
[11] S. Srivastava, R. Sinha, D. Roy, Toxicological effects of malachite green, Aquat. Toxicol., 66 (2004) 319–329.
[12] M.S. Gharavi-nakhjavani, A. Niazi, H. Hosseini, M. Aminzare, R. Dizaji, B. Tajdar-oranj, A. Mirza Alizadeh, Malachite green and leucomalachite green in fish: a global systematic review and meta-analysis, Environ. Sci. Pollut. Res., 30 (2023) 48911–48927.
[13] M.M. Ali, S. Nassar, N.A. Nafady, E.M. Mohamed, Bioremediation of malachite green dye toxicity under optimized conditions by Rhodotorula mucilaginosa AUMC13567, BMC Biotechnology, 25 (2025) 39.
[14] T. Das, A. Debnath, M.S. Manna, Adsorption of malachite green by Aegle marmelos-derived activated biochar: Novelty assessment through phytotoxicity tests and economic analysis, J. Indian Chem. Soc., 101 (2024) 101219.
[15] J.V.B. Borba, V.M. Alves, R.C. Braga, D.R. Korn, K. Overdahl, A.C. Silva, S.U.S. Hall, E. Overdahl, N. Kleinstreuer, J. Strickland, D. Allen, C.H. Andrade, E.N. Muratov, A. Tropsha, STopTox: An in Silico Alternative to Animal Testing for Acute Systemic and Topical Toxicity, Environ. Health Perspect., 130 (2022).
[16] J. Dong, N.-N. Wang, Z.-J. Yao, L. Zhang, Y. Cheng, D. Ouyang, A.-P. Lu, D.-S. Cao, ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database, Journal of Cheminformatics, 10 (2018) 29.
[17] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, X. Chen, T. Hou, D. Cao, ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Res., 49 (2021) W5–W14.
[18] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P.W. Lee, Y. Tang, admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties, Journal of Chemical Information and Modeling, 52 (2012) 3099–3105.
[19] Y. Gu, Z. Yu, Y. Wang, L. Chen, C. Lou, C. Yang, W. Li, G. Liu, Y. Tang, admetSAR3.0: a comprehensive platform for exploration, prediction and optimization of chemical ADMET properties, Nucleic Acids Res., 52 (2024) W432–W438.
[20] Y. Gu, C. Lou, Y. Tang, Chapter 14 - admetSAR—A valuable tool for assisting safety evaluation, in: H. Hong (Ed.) QSAR in Safety Evaluation and Risk Assessment, Academic Press2024, pp. 187–201.
[21] P. Banerjee, E. Kemmler, M. Dunkel, R. Preissner, ProTox 3.0: a webserver for the prediction of toxicity of chemicals, Nucleic Acids Res., 52 (2024) W513–W520.
[22] M.N. Drwal, P. Banerjee, M. Dunkel, M.R. Wettig, R. Preissner, ProTox: a web server for the in silico prediction of rodent oral toxicity, Nucleic Acids Res., 42 (2014) W53–W58.
[23] C.H. Chen, C.F. Chang, S.M. Liu, Partial degradation mechanisms of malachite green and methyl violet B by *Shewanella* decolorationis NTOU1 under anaerobic conditions, J. Hazard. Mater., 177 (2010) 281–289.
[24] M. Montazer, A. Bagheri Pebdeni, M.N. Sheikholeslami, S. Dehghan Abkenar, A. Firoozbakhtian, M. Hosseini, E.N. Dragoi, Synthesis of cuttlebone/ carbon quantum dots/nickel oxide nanocomposite for visible light photodegradation of

malachite green used for environmental remediation, Chemosphere, 333 (2023) 138880.
[25] S.-C. Wei, S. Fan, C.-W. Lien, B. Unnikrishnan, Y.-S. Wang, H.-W. Chu, C.-C. Huang, P.-H. Hsu, H.-T. Chang, Graphene oxide membrane as an efficient extraction and ionization substrate for spray-mass spectrometric analysis of malachite green and its metabolite in fish samples, Anal. Chim. Acta, 1003 (2018) 42–48.