ОБЗОР КОМПЬЮТЕРНЫХ ПРОГРАММ, ПРИМЕНЯЕМЫХ В БИОЛОГИЧЕСКИХ И ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

3. ИСПОЛЬЗОВАНИЕ ПРИКЛАДНОЙ ПРОГРАММЫ AXIS ДЛЯ ОЦЕНКИ КРУГОВОЙ СТАТИСТИКИ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

А.В. Мацюра, М.В. Мацюра
Мелитопольский государственный педагогический университет

В современной экологии достаточно широко распространены данные круговой статистики – распределение морских течений, направлений ветра, перемещения животных, как например, мигрирующих птиц и др. Большинство достаточно мощных статистических пакетов, широко применяемых современными экологами в нашей стране и за рубежом – такие, как SPSS, Statistica, Microcal Origin, SPLUS не позволяют выполнить грамотный анализ круговых данных, представить его графически и транспонировать полученные результаты в стандартные текстовые редакторы для дальнейшего использования.

Axis – программа, которая предлагает графические и аналитические методы, обычно используемые биологами, геологами и археологами для анализа круговых данных в операционной системе Windows. При помощи подобной программы разнообразные периодические данные могут быть представлены и проанализированы с использованием основных методов математической обработки круговой статистики (Fisher, 2003). Данные, которые использует программа: часы – время согласно 24-часовому периоду, дни – время согласно 365-дневному году, дни – время согласно 29,5-дневному лунному циклу, градусы – угол относительно 360° цикла, радианы – угол относительно 2π периода или цикла.

Методы и их обсуждение

1. Корреляция между выборками.

Эти методы представляют собой статистические тесты существования зависимости между двумя переменными. Подобные методы могут быть применены к набору начальных данных, который содержит минимум две сравниваемых выборки. Программа предлагает две статистических процедуры:

1) Т-линейная зависимость;

2) тест на внесение случайности.

Т-линейная зависимость аналогична простой линейной корреляции (Fisher, Lee, 1998, 2002). Чем ближе значение статистики к –1 или 1, тем больше степень отрицательной или положительной зависимости между двумя переменными. Статистическая достоверность зависимости отвергается, если значение не отличается значительно от ноля. Тест на случайность – это генеральный тест (Rothman, 1997) для гипотезы, что оба круговых распределения независимы.

2. Проверка на единообразие или произвольность.

Наблюдаемое распределение может быть проверено на единообразие, чтобы выяснить, одинаково ли вероятны все зафиксированные направления. Если необходимо проверить данные на любое отклонение от произвольного распределения, лучше использовать объемный тест. Однако, в силу того, что объемный тест способен обнаружить любое отклонение от произвольного, он не будет фиксировать специфические виды отклонения от равномерности. Например, для того, чтобы проверить, встречается ли одно специфическое направление более часто, чем ожидается при случайном распределении, то в этом случае более подходящим будет Rayleigh тест.

В распоряжении исследователя несколько статистических тестов: объемный тест для случайного распределения; Rayleigh тест на неопределенное среднее направление; Rayleigh тест на определенное среднее направление; тест Watson'а для одной выборки на равномерность; проверка медианного направления на указанное значение. Различные процедуры используются для несгруппированных и сгруппированных данных.

a) Объемный тест для сгруппированных данных. Этот тест используется, чтобы проверить, беспорядочно или однородно распределены наблюдения. Углы для n наблюдений сортируются от самого малого до наибольшего и каждый делится на 2π. В основу теста положена гипотеза, что выборка соответствует равномерному распределению (Arsham, 1998).

b) Объемный тест для сгруппированных данных. Данный тест применяется, чтобы определить, беспорядочно или однородно распределены наблюдения. Он используется для наблюдений, которые зарегистрированы с интервалом 5° или 10°. Число наблюдений в каждом интервале выверяется, используя Chi-squared тест. Если есть k интервалов в каждом цикле и n наблюдений, тогда в среднем должно быть $m=n/k$ наблюдений в интервале. Любое большое отклонение от этой средней величины указывает на отсутствие равномерности.

Y-тест рассчитывается по формуле:

$$Y = \left(k/n\right)\sum_{i=1}^{k} n_i^2 - n$$

где n_i – число наблюдения в i интервале.

Гипотеза на равномерность отвергается, если Y

превышает критическое значение для Chi-squared распределения с *k-1* степенями свободы.

c) Rayleigh тест – неопределенное среднее направление. Предоставляет *n* серии угловых измерений *u* и вычисляет:

$$C = \sum_{i=1}^{n} \cos(\theta_i)$$

$$S = \sum_{i=1}^{n} \sin(\theta_i)$$

где средний результирующий вектор:

$$\overline{R} = \frac{1}{n}\sqrt{C^2 + S^2}$$

Нуль-гипотеза на единообразие отвергается, если средний результирующий вектор слишком большой. Если данные сгруппированы, упомянутый выше тест применяется для каждого значения данных, вычисляя середину сгруппированного интервала.

d) Rayleigh тест – указанное среднее направление. В начале вычислении будет предложено выбрать данные и угол, с которым будет сравнен массив данных. В итоге получается *n* серий угловых измерений θ, вычисляя те же значения, что и предыдущий тест. Потом, используя среднее направление *m* и определенный пользователем угол θ вычисляется средний результирующий вектор:

$$\overline{R}_0 - \overline{R}\cos(\mu - \theta)$$

Нуль-гипотеза, которая предполагает, что нет никакой существенной разницы между *m* и θ, отвергается, если значение теста слишком большое.

e) Тест Ватсона на равномерность для одной выборки. Это – непараметрический тест на произвольность. Для начала *n* угловых наблюдений делятся на 360°, чтобы создать u_i преобразованных переменных. Затем рассчитывается статистический тест:

$$U^2 = \sum u_i^2 - \frac{\left(\sum u_i\right)^2}{n} - \frac{2}{n}\sum i u_i + (n+1)\overline{u} + \frac{n}{12}$$

Значение теста сравнивается с таблицей критических значений, чтобы определить, существует ли значительное отклонение от единообразия.

g) Проверка медианного направления на указанное значение. Необходимо выбрать массив данных и медианный угол, с которым массив будет сравнен. Пусть *m* – число наблюдений в дуге между указанным углом и указанным углом плюс 180°, которое не равно указанному углу. Если истинное медианное направление равно указанному углу тогда значение m не должно намного отличаться от $(n–k)/2$, где *n* – общее число наблюдений и *k* – число наблюдений, которое равняется указанному углу.

Формула теста:

$$Y = (2m - n + k)^2 \Big/ (n - k)$$

Статистическая гипотеза, что указанный угол равняется фактической медиане, отвергается, если *Y* превышает критическое значение для Chi-Squared теста с одной степенью свободы.

3. Сравнения между выборками.

3.1. Сравнение среднего направления.

3.1.1. P-метод.

Применение этого метода целесообразно, когда все образцы имеют одинаковую круговую дисперсию. Используя средние направления каждого из *r* образцов, вычисляется взвешенная средняя круговая дисперсия и Y-статистика. Нуль-гипотеза о том, что все выборки имеют одинаковое направление, отвергается, если превышает критическое значение для Chi-Squared распределения с *r–1* степенями свободы.

3.1.2. M-метод.

Этот метод применяется, когда выборки имеют различные круговые дисперсии. Используя средние направления каждого из *r* образцов вычисляется Y-статистика. Нуль-гипотеза, что все выборки имеют одинаковое направление, отвергается, если превышает критическое значение для Chi-Squared распределения с *r–1* степенями свободы.

3.2. Тест на общую медиану.

Для проведения теста необходимо иметь не менее 10 наблюдений в каждой выборке. Вычисления производятся по следующему алгоритму:

1) вычисляется срединное направление всех наблюдений *N*;

2) для каждого из r образцов, m_i – число, значений которого менее чем групповая медиана;

3) вычисляется $M = m_1 + ... m_r$;

4) вычисляется значение теста:

$$P_r = \left\{N^2 / [M(N-M)]\right\} \sum_{i=1}^{r} m_i^2 \Big/ n_i - NM \Big/ (N-M)$$

Гипотеза, что медианные срединные направления различных образцов одинаковы, отвергается, если P_r превышает критическое значение для Chi-Squared распределения с *r–1* степенями свободы.

4. Графическое представление данных.

Все графики и диаграммы могут быть сохранены в форматах (*.wmf), (*.bmp), (*.jpg). Программа предлагает большой набор графических средств для отображения данных. Данные могут быть представлены при помощи линейных гистограмм, которые позволяют отобразить их в линейном масштабе (Fisher, Powell, 2000).

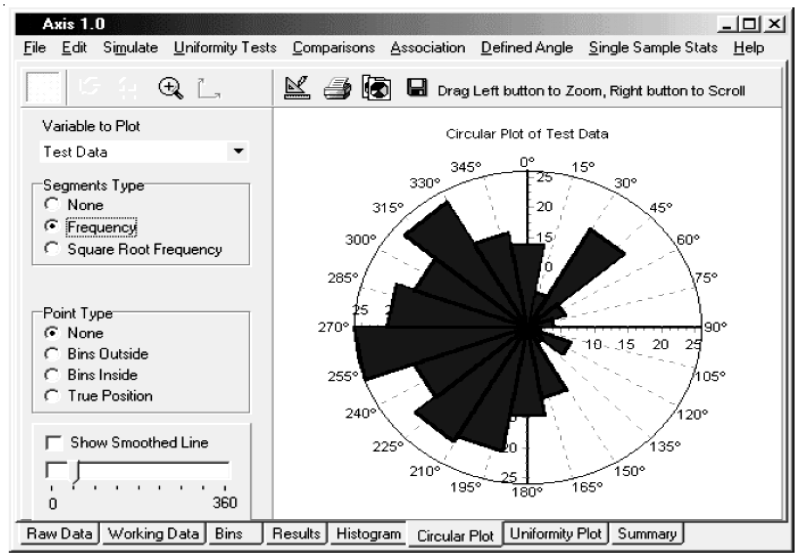
Вообще, круговые данные гораздо труднее анализировать, чем линейные – поэтому Axis предлагает ряд графиков с возможностью одновременно отобразить данные при помощи нескольких диаграмм, например, можно представить результаты наблюдений в виде простой круговой диаграммы, частотной гистограммы и сглаженной кривой.

В графическом наборе программы представлены: график неструктурированных данных, угловые гисто-

граммы, диаграммы направленности, графики сглаженных кривых. Для сглаживания кривых используется гармонический анализ.

На рисунке представлен скриншот графического анализа данных.

Обрабатывая данные, программа вычисляет главные элементы круговой статистики, среди которых: среднее направление, нижний и верхний предел 95% конфиденциального интервала, средний результирующий вектор, круговая дисперсия, круговое стандартное отклонение, медиана, нижний и верхний предел 95% конфиденциального интервала медианы, коэффициент асимметрии, коэффициент эксцесса.



Скриншот графического анализа данных – диаграмма направленности.

Выводы

Данная программа может быть с успехом применена для анализа круговых данных, например миграционных перемещений птиц, как в исследовательском, так и в учебном процессе. Основной трудностью, на наш взгляд, является англоязычный интерфейс программы, что несколько затрудняет область ее применения. Вместе с тем, простота и легкость вычислений, графического представления данных и аналитического блока позволят данной программе занять определенное место среди статистических пакетов, которые уже используются отечественными экологами при преподавании экологических дисциплин и анализе данных круговой статистики.

Литература

Arsham H. (1998): Kuiper's P-value as a measuring tool and decision procedure for the goodness-of-fit test. - J. Appl. Statist. 15: 131-135.

Fisher N.I. (2003): Statistical Analysis of Circular Data. Cambridge University Press. 1-277.

Fisher N.I., Lee A.J. (1998): A correlation coefficient for circular data. - Biometrika. 70: 327-332.

Fisher N.I., Lee A.J. (2002): Correlation coefficients for random variables on a unit sphere or hypersphere. - Biometrika. 73: 159-164.

Fisher N.I., Powell C.McA. (2000): Statistical analysis of two-dimensional palaeocurrent data: Methods and examples. - Aust. J. Earth Sci. 36: 91-107.

Rothman E.D. (1997): Tests for coordinate independence for bivariate sample on a torus. - Ann. Math. Statist. 42: 1962-1969.