УДК581.1.03.

***М.А.Живетьев 1, И.М. Романова 2***

**БИОИНФОРМАЦИОННЫЙ АНАЛИЗ В ЭКОЛОГИИ ХВОЙНЫХ И ТРАВЯНИСТЫХ РАСТЕНИЙ**

**Аннотация.** Целью данной работы стала практическая реализация статистических методов анализа с помощью среды программирования R и Statistica 7.0 для определения зависимости между абиотическими факторами и биотическими характеристиками растений. Проведены дисперсионный и корреляционный анализы, а также применены методы многомерной статистики для выявления влияния внешних условий на биохимический состав хвои сосны и листьев лекарственных растений пяти видов.

**Ключевые слова.** *Среда программирования R, Statistica, хвоя сосны, лекарственные растения.*

**Об авторах:**

*1 К.б.н., н.с., Сибирский институт физиологии и биохимии растений СО РАН, магистрант, Институт высоких технологий ИРНИТУ, г. Иркутск,* *nik.19@mail.ru*

*2 Ведущий инженер, Сибирский институт физиологии и биохимии растений СО РАН, Иркутск,* *irina170885@mail.ru*

**Введение**

Появление первых автоматических счетно-вычислительных машин, а позднее ЭВМ и персональных компьютеров, существенно расширили возможности в вычислительных операциях над большими массивами чисел. Параллельно шло совершенствование математического инструментария и методов статистического анализа. Результатом развития этих направлений стало то, что любой современный исследователь получил возможность привлекать широкий спектр универсальных программ для обработки и осмысления полученных им научных данных. Бурное развитие прикладной статистики за последние десятилетия привело к тому, что без ее применения не обходится ни одна качественная научная публикация во всех отраслях и дисциплинах, включая биологические науки. Все сильнее приходит осознание того, что без правильной статистической обработки данных невозможно получить научно-обоснованных выводов [1].

Не вызывает сомнения актуальность расширения применения математического и статистического инструментария для решения прикладных задач – в биологии и экологии, фармацевтике, медицинских и других науках [2]. Современные технологии обработки данных позволяют выявлять закономерности, которые человек самостоятельно сделать не в состоянии.

Ранее нами на основании полученных экспериментальных данных по хвойным и лекарственным растениям Прибайкальского региона были cделаны предварительные выводы, но без привлечения статистического и программного инструментария.

Еще в 1992 году А.И. Орлов утверждал [1], что без применения статистических методов научный потенциал используется далеко не в полной мере. Применение статистических методов как вид инженерной деятельности дает дополнительный стимул развития другим наукам, как это наблюдается на стыке любых дисциплин. В России до сих пор сохраняется разрыв между наукой о методах обработки данных и практикой их использования. Тем не менее, этот разрыв сокращается, в том числе за счет адаптации накопленных в прикладной математической статистике методов для нужд исследователей, например, через создание программных продуктов. Поэтому целью дальнейшей работы стала практическая реализация статистических методов анализа с помощью среды программирования R и Statistica 7.0 для определения зависимости между абиотическими факторами и биотическими характеристиками растений.

**1. Дисперсионный анализ влияния внешних условий на состав жирных кислот в хвое сосны обыкновенной**

Из абиотических факторов для анализа были выбраны: температура атмосферного воздуха и количество атмосферных осадков. Биохимические данные, полученные на базе СИФИБР СО РАН, включали содержания 32-х разных жирных кислот в хвое сосны сибирской, отобранной помесячно в разных географических точках Иркутской области – деревне Олха и поселке Мельничная Падь, заведомо отличающихся по температурам атмосферного воздуха и выпадению осадков. Дополнительным усложняющим аспектом было проведение отбора хвои четырех разных возрастов – от нуля (новая, молодая хвоя) до трех лет. Получившийся в итоге массив данных было проблематично осмыслить без привлечения статистических расчетов.

Метод дисперсионного анализа, или Analysis of Variance (ANOVA), был разработан Р. Фишером и используется для анализа связи между количественной зависимой переменной (Y) и качественными независимыми переменными или факторами (X1, …, Xn). В качестве таких факторов, не допускающих количественного описания, могут быть пол, биологический вид, год, тип почвы или вносимых удобрений и др. Конкретные значения фактора именуют его уровнями, количество которых должно быть не менее двух [3].

Дисперсионный анализ основывается на разложении (анализе) общей дисперсии (варьирования) зависимой переменной на компоненты, каждая из которых соответствует определенному источнику изменчивости измеряемой переменной [3].

Выделяют однофакторный дисперсионный анализ, который применяется для исследования влияния только одного фактора на количественную переменную. Результаты однофакторного дисперсионного анализа в общем виде представляются в виде сумм квадратов с указанием степеней свободы и последующего вычисления среднего квадрата и их отношений с указанием уровня значимости р [3].

В случаях, когда р < 0.05, можно делать обоснованный вывод, что фактор влияние оказывает.

На первом этапе осуществили однофакторый дисперсионный анализ по тесту Краскела-Уоллиса содержания жирных кислот попарно с такими факторами как место произрастания и возраст хвои (табл. 1).

Таблица 1.

**Результаты однофакторного дисперсионного анализа содержания жирных кислот в хвое сосны в зависимости от места произрастания деревьев и возраста хвои (по тесту Краскела-Уоллиса в R), значения p ˂ 0,05 помечены звездочкой**

|  |  |  |
| --- | --- | --- |
| Формула жирной кислоты | Р от места произрастания | Р от возраста хвои |
| С12:0 | 0,002\* | 0,097 |
| С12:0-i | 0,317 | 0,500 |
| С14:0 | 0,085 | 0,003\* |
| С15:0 | 0,148 | 0,054 |
| С15:0-a | 0,154 | 0,001\* |
| С16:0 | 0,344 | 0,984 |
| С16:1(n-5) | 0,146 | 0,002 |
| С16:1(n-7) | 0,071 | 0,486 |
| С16:1(n-9) | 0,602 | 0,374 |
| С16:1(n-11) | 0,317 | 0,500 |
| С16:2 | 0,154 | 0,217 |
| С16:2(n-6) | 0,079 | 0,135 |
| С17:0 | 0,352 | 0,065 |
| С16:3(n-3) | 0,158 | 0,826 |
| С18:0 | 0,867 | 0,536 |
| С18:1(n-9) | 0,171 | 0,004\* |
| С18:1(n-7) | 0,591 | 0,240 |
| С18:2(n-6) | 0,00002\* | 0,0168\* |
| С18:2(5,9) | 0,681 | 0,889 |
| С18:3(5,9,12) | 0,858 | 0,997 |
| С18:3(n-3) | 0,854 | 0,528 |
| С18:4(5,9,12,15) | 0,001\* | 0,484 |
| С20:0 | 0,008\* | 0,023\* |
| С20:1(n-9) | 0,558 | 0,512 |
| С20:2(n-9) | 0,240 | 0,214 |
| С20:3(5,11,14) | 0,092 | 0,011\* |
| С20:3(7,11,14) | 0,786 | 0,019\* |
| С20:3(11,14,17) | 0,031\* | 0,980 |
| С20:4(5,11,14,17) | 0,266 | 0,046\* |
| С20:4(5,8,11,14) | 0,079 | 0,9453 |
| С22:0 | 0,931 | 0,0014\* |

Для трех жирных кислот выявлена достоверная зависимость только от места произрастания, для семи – от возраста хвои, и для двух кислот – от обоих факторов. Причем наименьшие значения р были характерны для линолевой кислоты С18:2(n-6) – одной из ключевых жирных кислот всего растительного царства и преобладающей в тканях хвойных.

Результаты были визуализированы с использованием возможностей среды R и были представлены 64 рисунками, первая половина которых отражала зависимость содержания жирных кислот от места произрастания, а вторая – от возраста хвои.

Очевидно, такое разнообразие рисунков хорошо для конкретизации каждого частного случая, но громоздко и потому неудобно для осмысления совокупности данных. Поэтому далее была привлечена многомерная статистика.

**2. Многомерная статистика в изучении хвои сосны обыкновенной**

На втором этапе были взяты по возможности уже все известные параметры окружающей среды, морфометрические и биохимические данные по хвое в д. Олха и п. Мельничная Падь. Всего получилось 56 переменных, характеризующих каждую пробу: место отбора, возраст хвои, год отбора, месяц отбора, сумма осадков за месяц, температера среднемесячная, среднесуточная температура в день отбора проб, длина хвоинок, сырой вес хвои, сухой вес хвои, активность пероксидазы, шесть изоформ фермента пероксидаза, 32 жирные кислоты, сумма насыщенных (НЖК) и ненасыщенных (ННЖК) жирных кислот, коэффициент насыщенности (К) и инжекс двойных связей (ИДС) и активность ферментов, осуществляющих взаимопревращения жирных кислот (SDR, ODR и LDR).

Если в одномерной математической статистике результаты наблюдений сводятся к одному действительному числу, то в многомерном статистическом анализе результат наблюдения за объектом описывается несколькими числами, то есть вектором значений фиксированного числа чаще количественных, но иногда и качественных признаков, изученных у объекта [4].

На практике для осуществления математической статистики (как совокупности методов обработки данных) требуется привлекать компьютер и программы расчета, по возможности соответствующих современному научному уровню [4].

Расчеты по методу главных компонент, выполненные как в R, так и в Statistica 7.0, позволили выявить сильное влияние количества выпавших осадков на активность пероксидазы в тканях хвои сосны. Выраженного влияния температуры атмосферного воздуха на активность пероксидазы в течение года не прослеживалось. Причем среднемесячные температуры не влияли на активность фермента, а более высокие температуры воздуха непосредственно в день отбора проб оказывали незначительное положительное влияние на активность пероксидазы. В то же время, недостаточное выпадение осадков относительно сильно увеличивало активность этого фермента.

Так как увеличение активности пероксидазы считается маркером стрессового состояния, можно статистически обоснованно утверждать, что по всей видимости оводненность тканей для хвойных является более критичной, чем температура окружающей среды.

Ниже приведены нагрузки главных компонент для всего массива данных (табл. 2).

Нагрузки компонент являются вершинами векторов в двух или трехмерном пространстве. По этим числовым значениям можно построить в новых координатах главных компонент график сразу со всеми 56 точками, соответствующим исходным 56 параметрам внешней среды и биохимическим характеристикам. Для простоты восприятия, можно построить несколько графиков по отдельным группам параметров.

Для интерпретации результатов важно понимание того, что перпендикулярные вектора – это нулевая корреляция (нет зависимости), а разнонаправленные – обратная, а близость вершины векторов к началу координат свидетельствует о минимальной зависимости от других параметров и факторов. В результате можно видеть в одной системе координат и зависимые и независимые параметры.

Метод главных компонент позволил выявить ряд закономерностей. Так, год и месяц отбора имели близкое влияние, которое коррелировало с температурами (которые тоже почти совпадали, сливаясь в одну точку) и осадками. Таким образом, на погодные условия влияли год отбора и месяц, а не место произрастания сосны. Это понятно, потому что п. Мельничная Падь и д. Олха расположены относительно близко и метеорологические условия в этих точках не сильно отличались.

Наоборот, морфометрические показатели зависели от возраста хвои, что очевидно, и места произрастания, а не от времени года. Причем более информативным показателем, отражающим влияние точки сбора и возраста хвои, выступает длина хвоинок, а не ее вес. Однако, к сумме осадков был ближе всего сырой вес хвоинок.

Ближе всего к началу координат оказалась общая активность пероксидазы, активность которой несколько уменьшалась в ответ на увеличение выпадения осадков и повышение температуры. И наоборот, слабая положительная корреляция активности фермента наблюдается в ответ на увеличение возраста хвоинок и ростовые характеристики.

Таблица 2.

**Нагрузки трех первых главных компонент, рассчитанные в Statistica 7.0**

|  |  |
| --- | --- |
| Изученныепараметры | Главные компоненты |
| 1 | 2 | 3 |
| Место отбора | -0,32839 | 0,274442 | -0,540292 |
| Возраст хвои | -0,38832 | 0,141604 | 0,263034 |
| Год отбора | 0,08149 | -0,201429 | -0,465276 |
| Месяц отбора | 0,03239 | -0,285057 | 0,127400 |
| Сумма осадков за месяц | 0,05828 | -0,326146 | -0,454600 |
| Температера среднемесячная | 0,14708 | -0,537734 | -0,554270 |
| Средняя температура в день отбора проб | 0,15606 | -0,546726 | -0,499960 |
| Длина хвоинок | -0,39645 | 0,016154 | -0,500890 |
| Вес хвои (сырая масса) | -0,33182 | -0,033800 | -0,573737 |
| Вес хвои (сухая масса) | -0,35376 | 0,013055 | -0,542348 |
| Активность пероксидазы | -0,02371 | 0,099862 | -0,175617 |
| Изоформа 1 | 0,03818 | -0,176480 | -0,086683 |
| Изоформа 2 | -0,23937 | -0,073363 | 0,350456 |
| Изоформа 3 | -0,08679 | 0,143149 | 0,390526 |
| Изоформа 4 | -0,02023 | 0,161188 | 0,060058 |
| Изоформа 5 | 0,02022 | 0,233518 | 0,397025 |
| Изоформа 6 | -0,04489 | -0,276575 | 0,262244 |
| C12:0 | -0,02337 | -0,285674 | 0,368324 |
| C12:0-i | -0,05791 | -0,224110 | 0,057068 |
| C14:0 | -0,32825 | -0,160018 | 0,320444 |
| C15:0 | 0,75779 | -0,099444 | 0,010527 |
| C15:0-a | 0,68553 | -0,191931 | -0,167707 |
| C16:0 | -0,81643 | -0,308768 | -0,033954 |
| C16:1(n-5) | 0,92371 | -0,180209 | -0,187502 |
| C16:1(n-7) | 0,70674 | 0,296021 | -0,023174 |
| C16:1(n-9) | 0,60884 | -0,046569 | -0,239989 |
| C16:1(n-11) | 0,02410 | 0,096713 | -0,024522 |
| C16:2 | 0,37369 | 0,037697 | -0,080368 |
| C16:2(n-6) | 0,00242 | 0,529092 | 0,029380 |
| C17:0-a | -0,01590 | 0,032851 | 0,219882 |
| C17:0 | -0,18077 | -0,143453 | 0,013935 |
| C16:3(n-3) | 0,07147 | 0,172813 | 0,288968 |
| C18:0 | -0,42513 | -0,527503 | -0,110128 |
| C18:1(n-9) | -0,31736 | 0,607121 | -0,233005 |
| C18:1(n-7) | 0,66957 | 0,039266 | 0,019307 |
| C18:2(n-6) | -0,70059 | 0,466292 | -0,171800 |
| C18:2(5.9) | 0,43985 | 0,282073 | -0,021140 |
| C18:3(5.9.12) | 0,82973 | 0,110699 | 0,024427 |
| C18:3(n-3) | -0,78584 | -0,330810 | -0,009493 |
| C18:4(5,9,12,15) | 0,53896 | 0,243305 | 0,407978 |
| C20:0 | 0,92323 | -0,181779 | -0,113387 |
| C20:1(n-9) | 0,78372 | 0,205282 | -0,010138 |
| C20:2(n-9) | 0,04666 | 0,573998 | 0,143021 |
| C20:3(5.11.14) | -0,43196 | 0,458700 | 0,286307 |
| C20:3(7.11.14) | -0,01312 | 0,462625 | 0,181296 |
| C20:3(11.14.17) | 0,82805 | 0,000701 | 0,046839 |
| C20:4(5.11.14.17) | 0,26623 | 0,454402 | 0,222424 |
| C20:4(5.8.11.14) | 0,00409 | 0,142818 | 0,111665 |
| C22:0 | -0,33137 | 0,186181 | 0,513820 |
| НЖК | 0,04180 | -0,842251 | 0,353984 |
| ННЖК | -0,04037 | 0,831481 | -0,379921 |
| К | -0,04244 | 0,817998 | -0,375147 |
| ИДС | 0,03000 | 0,424802 | 0,127786 |
| SDR | 0,05814 | 0,784810 | 0,002118 |
| ODR | 0,11128 | -0,624411 | 0,197286 |
| LDR | -0,09235 | -0,619467 | 0,291584 |

Сильнее выражено влияние отдельных изоформ пероксидазы.

По расположению вершины вектора каждой из изоформ пероксидазы можно говорить с какими факторами и показателями каждая из них находится во взаимозависимости. Так, изоформа 2 по всей видимости связана с ростовыми процессами в хвое, а изоформа 1 – с увеличением температуры и обильными осадками. Изоформа 4 сильнее всего коррелирует с возрастом хвои.

Так как суммарная активность фермента складывается из работы каждой из изоформ, то неудивительно, что при разнонаправленной работе каждой из изоформ, общая активность пероксидазы относительно стабильна.

На жирнокислотный состав могут воздействовать изоформы пероксидазы под номерами 6, 3 и 5. Причем, если изоформа 6 коррелирует с насыщенными жирными кислотами и синтезом стеариновой и α-линоленовой жирных кислот и работой десатураз, выраженных индексами ODR и LDR, то изоформы 5 и 3, наоборот, способствуют синтезу ненасыщенных жирных кислот, увеличению индекса двойных связей и SDR.

Суммы ННЖК и НЖК находятся в достоверной сильной обратной зависимости у противоположных концов оси первой главной компоненты. Это логично связано с тем, что увеличение суммы ННЖК неминуемо приводило к уменьшению содержания НЖК, из которых те синтезируются:

 SDR ODR LDR

С18:0 -----------> С18:1 (n-9) -----------> С18:2 (n-6) -----------> С18:3 (n-3)

стеариновая олеиновая линолевая α-линоленовая

Более высокие температуры сопровождались увеличением суммы НЖК, LDR, ODR, содержания стеариновой (С18:0) и лауриновой (С12:0) кислот.

Обращает на себя внимание то, что другие (и большинство) насыщенных жирных кислот находятся «по экватору» графика, причем в основном ниже нуля по второй оси. Это физиологически обьяснимо: они вносят меньший вклад в содержание НЖК хвои, при этом имея более высокие температуры плавления и межфазового перехода, поэтому коррелируют с более высокими температурами.

Наоборот, ненасыщенные жирные кислоты в подавляющем большинстве расположены в верхней и «экваториальной» части рисунка.

С17а занимает самый центр графика и является самой ни с чем не зависимой, что не удивительно, учитывая ее следовые количества в пробах.

Наличие в центре (вблизи координаты (0,0)) скопления большинства данных может быть связано, во-первых, с тем, что они являются малозависимыми, а во-вторых может быть обусловлено низкой дисперсией главных компонент в данном конкретном анализе.

**3. Корреляционный анализ применительно к лекарственным растениям Прибайкалья**

Целью любой науки является выявление связей и закономерностей между признаками или характеристиками окружающего мира – будь то язык, культура, живая или неживая материя. Соответственно для изучения интересующей исследователя зависимости или закономерности между двумя характеристиками требуется доказать наличие связи между ними и изучить эту связь. Для этого чаще всего в экологии используют корреляционный анализ [5]. Этот статистический метод относительно прост и легко применим на практике, чем обусловлено предпочтение в его использовании. При нормальном распределении обоих сравниваемых параметров статистические выводы базируют на выборочном коэффициенте линейной корреляции, по Пирсону. Если оба или один из параметров распределен по ненормальному закону, то используются коэффициенты ранговой корреляции Кендалла или Спирмена, а для дискретных качественных признаков – критерий хи-квадрат [5].

В работе использовались метеорологические сведения о температуре атмосферного воздуха в Иркутске и Иркутской области и научно-исследовательские данные по содержанию водорастворимых сахаров в листьях лекарственных растений пяти видов: одуванчик лекарственный, подорожник большой, тысячелистник азиатский, манжетка городковатая и вероника дубравная. Для выявления взаимосвязи этих двух параметров использовались возможности среды программирования R для расчета корреляции.

Результаты проведенного теста Шапиро-Уилка на нормальность распределения выборки и рассчитанные ею коэффициенты корреляции сведены в таблицу 3.

Таблица 3.

**Результаты вычисления вероятности принятия нулевой гипотезы и корреляции среднесуточных температур с содержанием водорастворимых сахаров в листьях растений пяти биологических видов**

|  |  |  |
| --- | --- | --- |
| Вид растения | Тест на нормальность распределения Шапиро-Уилка | Коэффициенты корреляции |
| Среднесуточная температура | Концентрация сахаров | по Пирсону | по Спирману |
| W | Р | W | Р | Cor | Rho |
| Одуванчик | 0.83730 | 0.04096 | 0.87822 | 0.12450\* | -0.834898 | -0.7090909\* |
| Подорожник | 0.83730 | 0.04096 | 0.84389 | 0.04911 | -0.8570318\* | -0.5636364 |
| Тысячелистник | 0.83730 | 0.04096 | 0.88340 | 0.14270\* | -0.8488697 | -0.6727273\* |
| Манжетка | 0.82378 | 0.12480\* | 0.83888 | 0.16190\* | -0.9210742 | -0.9000000\* |
| Вероника | 0.82378 | 0.12480\* | 0.84474 | 0.17840\* | -0.984773 | -0.6000000\* |

В случаях, когда хотя бы один из р >0,05 (в таблице помечены звездочкой «\*»), мы смотрим коэффициент корреляции по Спирману (тоже помечен), иначе – коэффициент корреляции по Пирсону (для подорожника).

В целом, на основании проведенного статистического анализа можно утверждать, что во всех случаях наблюдается обратная корреляция температуры и содержания сахаров в листьях растений, причем довольно высокая – от 60% у вероники и 67,3 у тысячелистника, до 85,7 у подорожника и 90% для манжетки. Иными словами, минус перед значением соответствующего коэффициента свидетельствует о том, что при понижении температуры воздуха происходит увеличение содержания сахаров в листьях растений. Меньше всего эта закономерность прослеживается у тысячелистника и вероники, а сильнее у подорожника и манжетки.

Выявленная закономерность имеет физиологическое объяснение. Сахар в тканях растений, в частности в цитоплазме и межклетниках, препятствует образованию льда при понижениях температуры, тем самым спасая растения от гибели при низких температурах.

В дальнейшем планируется показать другие возможности R-программирования для обсчета биологических данных с целью математически обоснованного выявления сложных взаимосвязей растительного организма с окружающей средой.

# Список использованных источников

1. Орлов А. И. О современных проблемах внедрения прикладной статистики и других статистических методов (обобщающая статья) / А. И. Орлов // Заводская лаборатория. – 1992. – №1. – С. 67–74.

2. Смагунова О. Н. Методы математической статистики в аналитической химии: учебное пособие / О. Н. Смагунова, О. М. Карпукова. – Ростов н/Д: Феникс, 2012. – 346 с.

3. Мятлеев В. Д. Теория вероятностей и математическая статистика. Математические модели: учебное пособие для студ. высш. учеб. заведений (Университетский учебник. Высшая математика и ее приложения к биологии) / В. Д. Мятлеев, Л. А. Панченко, Г. Ю. Ризниченко, А. Т. Терехин. – М.: Издательский центр «Академия», 2009. – 302 с.

4. Орлов А. И. Вероятность и прикладная статистика: основные факты: справочник. / А. И. Орлов. – М.: КНОРУС, 2010. –192 с.

5. Шипунов А. Б. Наглядная статистика: используем R! / А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников, С. А. Назарова, С. В. Петров, В. Г. Суфиянов. – М.: ДМК Пресс, 2014. – 298 с.

**M. A. Zhivetyev, I. M. Romanova**

**Bio-ecological analysis of conifers and herbaceous plants**

**Abstract.** *The aim of this work was the practical implementation of statistical methods of analysis using the R programming and Statistica 7.0 to define dependencies between abiotic factors and biotic characteristics of plants. Correlation analyses of variance were conducted, as well as applied multivariate statistical methods to detect the influence of external conditions on the biochemical composition of the pine needles and leaves of five species of medicinal plants.*

**Keywords.** *R programming, Statistica, pine needles, herbs.*