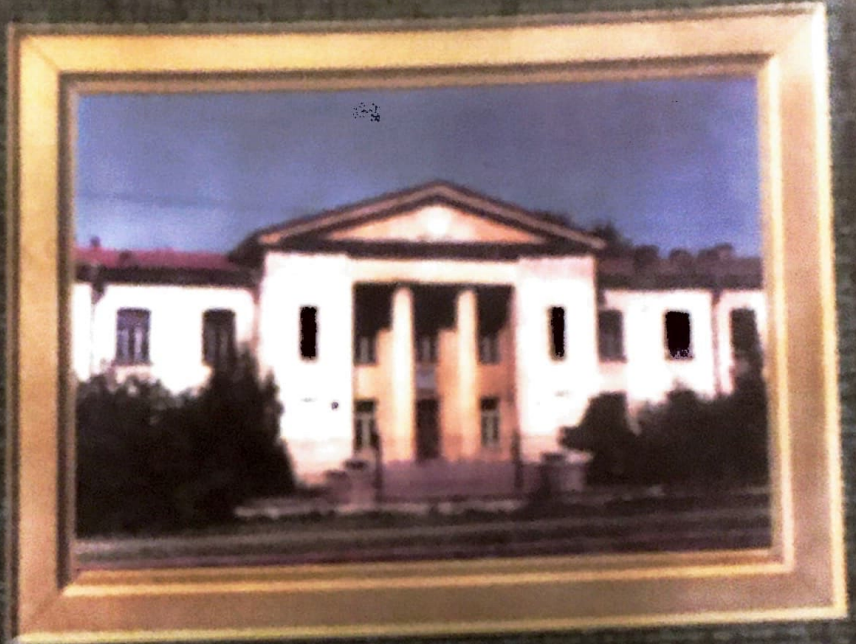




1932-2002



АГРОФИЗИЧЕСКИЙ
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ
ИНСТИТУТ

70 лет

СТАТИСТИЧЕСКИЕ МЕТОДЫ В АГРОФИЗИКЕ

В.П.Якушев, В.М.Буре, Т.М.Брунова

В науках, основывающих свои выводы на общепринятых приемах научного исследования – наблюдении и эксперименте, давно замечено, что нельзя ограничиваться лишь детализированными функциональными зависимостями, характеризующими изменчивость и динамику изучаемых показателей (признаков). Даже в лабораторных условиях при фиксации всех основных факторов, оказывающих воздействие на изучаемое явление, обнаруживается некоторая изменчивость в результатах опыта,

которая, как правило, в принципе не может быть устранена. Для описания подобной изменчивости используют идею случайности. В соответствии с этой идеей, результат определяется некоторым случайным событием, иначе говоря, природа словно бы бросает «кости», для осуществления окончательного выбора. Вопрос о том, существует ли случайность в природе, носит дискуссионный философский характер. В практической деятельности достаточно использовать идею случайности лишь как модель для описания непредсказуемой изменчивости, считая, что возникновение этой изменчивости обусловлено влиянием некоторых «неучтенных» факторов. Все вышесказанное в полной мере относится к экспериментальной агрофизике, как к экспериментальной науке, основанной, прежде всего на опытных исследованиях. Конечно, признание наличия случайной изменчивости ни в малейшей степени не означает отрицание существования закономерной причинно-обусловленной изменчивости, изучение которой составляет главную суть любой экспериментальной науки. Статистические методы в исследованиях АФИ использовались довольно широко. Следует отметить, что в институте О.М.Калининым был разработан первый в СССР пакет машинных программ по статистике. Полученные результаты имеют как самостоятельное значение для изучения статистических закономерностей изменчивости агрометеопараметров и взаимосвязи их между собой, так и прикладное – для решения различных задач агротехники и сельского хозяйства. Так, например, если задан вид культуры, известна почва и определена агротехника, то урожай зависит от случайных метеорологических факторов. Число проводимых наблюдений и разнообразных опытов в природных условиях (на полях) определяется требованием достаточности статистики, к тому же дело осложняется статистической нестационарностью, ибо процесс осложняется нелинейностью некоторого оператора, выявленного в процессе опытов или данных, получаемых на основе статистических наблюдений в полевых условиях. Даже в условиях искусственного климата необходимо вероятностное описание, ибо наблюдается разброс показателей у растений одного вида, сорта, возраста, в том числе и по интегральному показателю роста и развития – урожаю, даже развитие плодов на одном растении и зерен в одном колосе различны. Было показано, что средние во времени не равны среднему по множеству. Максимизация урожая носит статистический характер. При построении статистических моделей некоторых параметров почвы и приземного воздуха, которые рассматривались как нестационарные случайные процессы, проводилось детальное исследование суточной и годовой изменчивости их основных вероятностных характеристик, обсуждалась проблема получения их пространственно-

репрезентативных характеристик (Е.Е.Жуковский, И.И.Косенко, С.М.Мандельштам, А.Ф.Чудновский 1969; Е.Е.Жуковский, Т.Л.Киселева, С.М.Мандельштам, 1976; Е.Е.Жуковский, М.Г.Саноян, 1977; М.В.Петрова и др.).

Суть статистического подхода в экспериментальных науках заключается в разделении наблюдаемой изменчивости на две составляющие — закономерную составляющую и случайную, и в последующем изучении закономерной составляющей на фоне случайной. Еще один важный аспект понятия случайности связан с тем, что случайность способна породить закономерность. В статистической физике невозможно описать движение каждой отдельной молекулы газа в виду того, что оно носит хаотический случайный характер, однако взятые в совокупности молекулы газа ведут себя закономерно, подчиняясь известным законам физики. Важную роль играет учет случайной изменчивости в биологических исследованиях, где сложность исследуемых объектов особенно велика, что порождает возможность возникновения многочисленных связей и взаимодействий.

Нельзя анализировать опытные данные в агрофизике, игнорируя наличие в них случайной составляющей, другими словами, необходимо понимать, что любые математические модели являются лишь приближением к реальной действительности, и поэтому, кажущаяся «высокая точность» для многих моделей, на самом деле, оказывается иллюзорной при расширении сферы их применения, поскольку важную роль начинают играть неучтенные факторы или новые взаимодействия старых факторов. Особенно плохо, когда «высокая точность» достигается за счет увеличения сложности модели, поскольку сложная динамика, порождаемая усложненной моделью, может хорошо соответствовать ограниченному набору данных, полученному в эксперименте, но плохо соответствовать реальности при последующем использовании, именно за счет своей излишней сложности.

Далее рассмотрим различные аспекты статистического анализа экспериментальных данных с учетом присутствия случайной изменчивости в результатах наблюдений. Статистический анализ накопленных опытных данных можно классифицировать, используя различные подходы. Один из подходов заключается в разделении статистического анализа на два этапа: разведочный и подтверждающий. Разведочному или первичному анализу данных посвящена книга известного специалиста по математической статистике Дж.Тьюки. Первичный анализ, главным образом включает преобразования данных, а также различные способы их

наглядного представления, позволяющие выявить внутренние закономерности, проявляющиеся в данных.

Оценка погрешности интерполяции температуры приземного слоя воздуха в зависимости от вида интерполяции: ступенчатая, линейная, по условному математическому ожиданию с учетом одного или двух последовательных отсчетов — для восстановления непрерывной функции во времени по дискретной последовательности (Н.Г.Голант, Т.Л.Киселева, С.М.Мандельштам, А.Ф.Чудновский, 1968) показала, что линейная интерполяция предпочтительнее при восстановлении температурных кривых. На примере статистического анализа динамики пыльных бурь построена методика (И.Б.Усков, 1984) определения границ регионов, потенциально предрасположенных к возникновению однотипных экстремальных метеорологических явлений, и показана возможность применения теории распознавания образов к количественному описанию скачкообразного перехода системы почва-атмосфера в новое качественное состояние, а также априорной дифференциации по структуре вероятностных оценок пыльных бурь на местные и синоптически обусловленные.

В современных условиях при использовании статистических пакетов для персональных компьютеров (Ю.Н.Тюрин, А.А.Макаров, 1998) решению этих задач способствуют разнообразные методы описательной статистики, многочисленные графики и диаграммы. На втором этапе статистического анализа применяются различные методы оценки параметров и проверки статистических гипотез.

При статистическом анализе опытных данных возникает необходимость в формулировке и экспериментальной проверке некоторых предположительных утверждений (статистических гипотез) относительно стохастических систем, порождающих экспериментальные данные. Суть проверки заключается в количественной оценке степени согласованности экспериментальных данных с гипотезой. Проверка такой согласованности позволяет избежать неверных выводов, основанных на субъективном мнении исследователя, дает возможность объективной оценки полученных результатов. В этом смысле проверка статистических гипотез является важнейшим этапом статистического исследования. Можно выделить несколько основных типов статистических гипотез (С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин, 1983): 1) гипотезы о типе закона распределения исследуемой случайной величины; 2) гипотезы об однородности двух или нескольких обрабатываемых или некоторых характеристик обрабатываемых совокупностей; 3) гипотезы о числовых значениях параметров исследуемой генеральной совокупности; 4) гипотезы о типе зависимости между компонентами исследуемого многомер-

ного признака или гипотезы независимости компонент; 5) гипотезы независимости и стационарности обрабатываемого ряда наблюдений.

Гипотезы могут быть простыми и сложными. Простые гипотезы — однозначно определяют закон распределения генеральной совокупности. В противном случае, она называется сложной. Особо следует выделить гипотезы второго и четвертого типа. По существу, дисперсионный анализ предназначен для проверки некоторых частных гипотез второго типа. Гипотезы четвертого типа играют особенно важную роль в агрофизических исследованиях, поскольку цель конкретных исследований часто заключается в изучении различного вида зависимостей между признаками.

Другая часто возникающая задача в экспериментальных исследованиях — оценка параметров вероятностных распределений. Законы распределения, используемые для описания случайной изменчивости, содержат неизвестные параметры — среднее значение, дисперсию, коэффициент корреляции, вклады факторов, коэффициенты функциональных зависимостей и т.д. Исследователя обычно интересуют либо сами параметры, либо функции от них. По опытным данным можно оценить указанные величины, то есть приближенные значения, используя имеющиеся экспериментальные данные. Вопросам оценивания посвящена обширная литература, включающая как теоретические исследования свойств полученных оценок, так и построение различных вычислительных процедур оценивания. К задаче оценивания неизвестных параметров вероятностных распределений примыкает задача оценки статистических зависимостей, называемая регрессионным анализом. Методы линейного и нелинейного регрессионного анализа широко используются в опытных исследованиях. В решении практических задач оценки зависимостей возникает много трудных проблем, для решения которых необходимо глубокое понимание природы решаемой задачи, а также хорошее знание разнообразных методов оценки зависимостей и условий их применимости. Литература по статистическим оценкам зависимостей изобилует примерами неудачного, а порой ошибочного применения тех или иных способов построения регрессионных зависимостей. Можно сказать, что регрессионный анализ является одним из основных методов статистического анализа, однако следует иметь в виду, что положительный результат его применения зависит от тщательного исследования, в том числе адекватности полученных моделей. Проведение подобного изучения гораздо сложнее, чем формальное проведение хорошо известных вычислений по методу наименьших квадратов для линейной по неизвестным параметрам модели. Без основательного изучения адекватности модели, ее использование и интерпретация приводит к грубейшим ошибкам, полно-

стью искажающим содержание и смысл, при анализе опытных данных можно выделить группу методов, объединяемых названием – многомерный статистический анализ. Сюда, прежде всего, относятся факторный, дискриминантный и кластерный анализы, а также многомерное шкалирование. Необходимость использования всех этих методов, а также некоторых других, входящих в группу приемов многомерного статистического анализа, обусловлена, прежде всего, тем, что в одном эксперименте – например, при исследовании одного объекта, – измеряются сразу несколько характеристик, причем их число может быть довольно большим. В таких опытах каждое наблюдение представляется не одним единственным числом, а некоторым конечным набором чисел, в котором в определенном порядке записаны все измеренные характеристики объекта. Подобная ситуация часто возникает в практических исследованиях. Однако в настоящее время довольно хорошо разработана теория многомерной статистики лишь для многомерных гауссовских наблюдений, то есть когда генеральная совокупность имеет многомерное нормальное распределение. Но если распределение генеральной совокупности другое, то возникает сложность с интерпретацией получаемых в результате статистического анализа результатов. Ниже будет говориться о том, что в одномерном случае имеется альтернатива для параметрических методов – непараметрическая статистика, а вот в многомерной статистике, к сожалению, непараметрические методы, не имеют многомерного анализа. Тем не менее, практическое использование перечисленных выше многомерных методов показало их работоспособность и эффективность, хотя и в этой области статистического анализа имеется не мало примеров ошибочного применения многомерных статистических процедур. Остановимся на краткой характеристике перечисленных выше многомерных методов. При изучении сложных биологических объектов число и содержательный смысл факторов, определяющих свойства этих объектов, могут быть неизвестными исследователю. Для измерений могут быть доступны другие величины – признаки, зависящие от неизвестных факторов. Количество измеряемых признаков может быть большим, а количество факторов, оказывающих значимое влияние, может быть существенно меньше. Методы факторного анализа позволяют по имеющимся опытным данным оценить неизвестные факторы с помощью линейных комбинаций исходных признаков, а затем, применяя некоторые преобразования, получить интерпретации этих факторов.

Методы дискриминантного анализа предназначены для решения задачи классификации некоторых объектов по количественным характеристикам. Обычно предполагается, что объекты разбиты на несколько