МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет радиофизики и компьютерных технологий

Кафедра физики и аэрокосмических технологий

АЛГОРИТМЫ ВЫДЕЛЕНИЯ ОДНОРОДНЫХ УЧАСТКОВ ПО УРОЖАЙНОСТИ В ТОЧНОМ ЗЕМЛЕДЕЛИИ

Курсовая работа студента 4 курса
Иванова Алексея Леонидовича

Руководитель: старший преподаватель
кафедры физики и аэрокосмических технологий
ВЕРХОТУРОВА Е.В.

Минск, 2011

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Сформированная в 90-х годах прошлого столетия методология точного земледелия (ТЗ) является новым направлением в агрономической науке. Появление этого направления было обусловлено прогрессом в области компьютерной техники и информационных технологий, развитием средств космической связи, появлением рабочих органов сельскохозяйственных машин, способных осуществлять дифференцированное (в пределах поля) управление технологическими операциями.

Очевидно, что традиционный подход к ведению сельского хозяйства далеко не всегда эффективен и рационален. Неконтролируемое, часто необоснованное, увеличение норм вносимых минеральных удобрений не обеспечивает ожидаемого прироста урожая, кроме этого резко возрастают затраты невосполнимой энергии (нефти, газа, угля) на производство единицы продукции, в угрожающих размерах увеличивается загрязнение окружающей среды.

Существует ряд вопросов, на которые традиционный подход ответить не может: какой оптимальный урожай может быть получен в конкретных почвенно-климатических условиях, какие при этом следует привлечь ресурсы, как эффективно управлять формированием урожая, какие изменения будут происходить с почвой при применении той или иной технологии? Дать ответы на эти вопросы можно только при использовании качественно нового подхода, основанного на возможности выбора оптимального решения, в результате системного анализа имеющегося агроэкологического комплекса.

Таким качественно новым подходом является точное земледелие, которое позволяет решить задачу получения оптимального урожая путём дифференцированного применения агротехнологии в соответствии со складывающимися метеорологическими условиями, биолого-почвенными характеристиками пашни и возможностями хозяйства.

Согласно концепции точного земледелия любое поле представляет собой совокупность неоднородных участков, которые требуют различную степень обработки для получения наибольшей урожайности. Основным критерием для разделения поля на неоднородные участки служит средняя урожайность, получаемая с этих участков на протяжении наблюдаемого периода лет. Таким образом, появляется один из основных вопросов точ-

ного земледелия: поиск алгоритмов разделения неоднородного поля на однородные участки.

Поиск, реализация и сравнение алгоритмов выделения однородных по урожайности участков на неоднородном сельскохозяйственном поле и является целью моей курсовой работы.

1. КОНЦЕПЦИЯ И КОРНИ ТЕРМИНОЛОГИИ ТЗ

Термин «точное земледелие» появился в 90-е годы XX столетия как естественное понятие устойчивого земледелия. Процесс получения растениеводческой продукции распределён во времени и пространстве, т.е. осуществляется на некоторой территории и на определенной глубине почвенного покрова. Эта территория не является однородной даже в пределах одного поля или его части. По этой причине технологические операции, производимые на поле, должны быть дифференцированы не только во времени (с учётом изменчивости погодных условий) и по полям севооборота, но варьироваться также и в пределах одного поля [8].

Принципиальное отличие новой концепции заключается в том, что технология точного земледелия рассматривает каждое сельскохозяйственное поле как неоднородное. Оно разделяется на некоторое количество новых единиц управления, которые являются однородными (квазиоднородными) участками. Суть в том, что для получения с данного поля максимального количества качественной продукции для всех растений этого массива создаются оптимальные условия произрастания с учётом выявленной неоднородности участка. Такой подход, безусловно, не является новым. Если задаться вопросом, какой наименьший участок можно взять, чтобы создать оптимальные условия для растений, то это будет участок с отдельно взятым растением. Так и поступали, очевидно, первые земледельцы, когда сажали и удобряли растения вручную. При современном же крупномасштабном производстве стремиться к подобной цели можно, лишь призвав на помощь новые уникальные технологии.

Концепция повышения эффективности сельскохозяйственного производства на основе учёта пространственной и временной изменчивости параметров плодородия почв и состояния растений лежит в основе ресурсосберегающего и экологически безопасного производства и является перспективным направлением в агрономии. Появление новой технологии было обусловлено возросшими требованиями экологической безопасности земледелия и экономии удобрений и средств защиты растений, а также невозобновляемых ресурсов – горюче-смазочных материалов. Отметим, что разработка методологии точного земледелия не является «революционным скачком» в совершенствовании агротехнологии. Напротив, это естественный следующий шаг в агрономических исследова-

ниях, который должен учитывать все ранее полученные результаты в этом направлении, включая результаты исследований по разработке систем ландшафтного земледелия и адаптивного растениеводства.

В отечественной литературе это новое направление имеет два названия – «точное земледелие» и «координатное земледелие». Оба они имеют право на существование, хотя обозначают разные понятия [8].

Точное земледелие – фундаментальная наука, занимающаяся разработкой стратегии и тактики земледелия, а также оперативного управления продукционным процессом сельскохозяйственных растений с учётом биологических особенностей культуры и сорта, локальных условий почвенного питания растений и микроклиматических особенностей территории.

Координатное земледелие – прикладная наука, разрабатывающая дифференцированные технологии земледелия, направленные на получение заданных экономически и экологически обусловленных урожаев при максимальной экономии невозобновляемых ресурсов с учётом неоднородности почвенного покрова в пределах одного поля.

В англоязычной литературе также существуют два термина «precision agriculture» и «precision farming», которые можно считать аналогами двух понятий, обозначенных выше. Координатное земледелие является частью точного земледелия; такое разделение, однако, оправдано тем, что прикладные разработки имеют свою важную специфику. Например, конструирование рабочих органов сельхозмашин, управляемых бортовым компьютером, представляет собой важную практическую задачу, решаемую не только в рамках научно-исследовательских работ, но и посредством опытно-конструкторских изысканий.

Национальный исследовательский комитет США (US National Research Council) определяет понятие точного земледелия следующим образом: «Точное земледелие – стратегия управления, которая использует информационные технологии, извлекая данные из множественных источников с тем, чтобы принимать решения по управлению посевами».

Ключевыми словами в этом определении являются «управление посевами», «информационные технологии» и «использование данных из множественных источников».

Именно развитие информационных технологий открыло пути существенного совершенствования методов принятия решений в агрономии. Новые технологии, которые обусловили возможность перехода к концепции точного земледелия, связаны с появлением географических информационных систем, возможностью использования глобальной системы позиционирования с непосредственным вводом информации в бортовой компьютер, обеспечивающий управление механизмом, проводящим в поле ту или иную операцию. Стоит подчеркнуть, что решающую роль в этом процессе играет информационное обеспечение принятия управленческих решений – моделей, баз данных и знаний, экспертных систем, специальных программ.

2. ПОДСИСТЕМЫ ТОЧНОГО ЗЕМЛЕДЕЛИЯ

Развитие методологии точного земледелия (ТЗ) стало возможным благодаря беспрецедентному прорыву в разработке специальной техники и информационных технологий, которые были успешно интегрированы в сельское хозяйство. Внедрение ТЗ в сельскохозяйственную практику требует оснащения пользователей специальным оборудованием и программным обеспечением. Рассмотрим основные подсистемы, входящие как элементы в технологию точного земледелия.

2.1. Навигационная подсистема

Навигационная система – глобальная система позиционирования с вводом данных в бортовой компьютер. Именно с появлением ГСП открылась принципиальная возможность для перехода от традиционной технологии к координатному земледелию, при котором можно влиять на агроэкосистему с учётом локальной изменчивости почвенного покрова поля.

ГСП используется для определения координат мобильной сельскохозяйственной техники в поле. В настоящее время на нашей территории функционируют две системы глобального позиционирования: американская NAVSTAR и российская ГЛОНАСС. Они позволяют неограниченному числу объектов, имеющих приёмную аппаратуру, в режиме реального времени и с высокой точностью определять в любой точке планеты своё местоположение, скорость движения и ряд других параметров. Наибольшее распространение в ТЗ получила приемная аппаратура американской системы в связи с хорошо налаженным производством и полностью развернутой группировкой космических аппаратов. Российская система пока не имеет достаточного количества спутников на орбите, и поэтому не может обеспечить хорошей точности. Однако сейчас идет поэтапная модернизация системы ГЛОНАСС, и будем надеяться, что в ближайшем будущем качество сигнала улучшится. Кстати, применение аппаратуры, использующей сигналы обеих систем, позволяет с большей надежностью и намного эффективнее определять местоположение объектов.

Задача определения координат мобильной сельскохозяйственной техники является одной из главных в точном земледелии по двум причинам. Во-первых, таким путем решается задача дифференциации управления в пределах поля, и участки неоднородности

могут быть четко идентифицированы. Вторая причина заключается в том, что размеры управляющих воздействий, например, внесения удобрений, высева семян, обработки средствами защиты растений, должны варьироваться с учётом выявленной неоднородности по заданной технологической карте в режиме «on-line».

Навигационная система, устанавливаемая на сельскохозяйственной технике, включает в себя так называемый GPS-приёмник и бортовой компьютер с программным обеспечением. Этот комплекс позволяет вести запись текущих координат данного агрегата, высоты и других параметров с любыми заданными интервалами времени. Запись навигационных данных производится в широко известных форматах ESRI Shapefile и MapInfo, что позволяет легко импортировать их в любую ГИС для дальнейшей обработки и производства необходимых агротехнических расчётов.

При использовании технологии ТЗ применяются несколько видов приёмников с разным уровнем точности определения местоположения и, соответственно, различной ценовой категории. Для специалиста-агронома, например, в повседневной деятельности оптимален карманный портативный компьютер (КПК), сопряженный с GPS-приёмником. При стоимости 300-400 долларов эта связка позволяет выполнять много функций, вплоть до передачи на центральный пункт управления сведений о состоянии посевов и т.п.

Точность определения с помощью таких приборов составляет 10-15 метров, и для ТЗ такая точность недостаточна. В этом случае используют дополнительно принимаемую поправку (DGPS). Её можно получать различными способами: через глобальную платную спутниковую систему, европейскую бесплатную спутниковую систему, от морских радиомаяков и от локальных базовых станций. Платная годовая подписка (Omnistar, Rakal) при точности определения местоположения до одного метра стоит около 1000 долларов. Услуги Европейской системы EGNOS предоставляются бесплатно при точности 1-3 метра (стоимость GPS-приёмника с поддержкой EGNOS обойдётся потребителю приблизительно в 500 долларов). Стоимость локальной базовой станции составляет около 4000 долларов, при этом точность может достигать нескольких сантиметров. В зоне действия морских радиомаяков точность определения положения обычно бывает менее метра.

В настоящее время налажен выпуск двухчастотных GPS-приемников, с помощью которых возможно построение достаточно точных цифровых моделей рельефа местности. Однако такой тип GPS-приемников дорог и довольно неудобен для применения должным образом в фермерской практике. Имеются результаты изучения оценки возможности использования для этих целей более дешевых одночастотных плоских (горизонтальных) GPS-приемников. В частности, показано, что использование одночастотных GPS-приемников для построения цифровых моделей рельефа сельскохозяйственных полей целесообразно лишь в том случае, когда точность меньше метра.

Таким образом, потребитель может выбрать из всего спектра навигационных средств то оборудование, которое позволит ему наиболее эффективно осваивать технологию ТЗ.

2.2. Технические средства для получения и обработки информации

Получение информации о почвенном покрове, состоянии растений и их урожайности, степени поражения вредителями, болезнями и сорняками требует наличия соответствующих приборов и оборудования, снабженных методическими указаниями по их эксплуатации, а также специальных информационно-измерительных технологий [3]. Эффективность точного земледелия во многом зависит от того, как быстро и точно будут измерены те или иные параметры, характеризующие состояние агроценоза. Частота измерений (пространственная и временная) зависит от того, какова изменчивость измеряемого показателя. В связи с этим возникает большая необходимость в разработке специальных технических средств, используемых для автоматизированного сбора и анализа информации, характеризующей систему «почва – растения – деятельный слой атмосферы» [6].

Использование датчиков позволяет получать данные с гораздо большей разрешающей способностью, чем это возможно при использовании традиционных методов, когда отбираемые образцы почвы и растений анализируются в лаборатории [2, 4, 5]. Вместе с тем, разработка датчиков для оценки состояния среды обитания растений сталкивается с разного рода препятствиями, и по этой причине это направление значительно отстаёт от других технологий сбора данных.

Существует две наиболее приоритетные темы в развитии этой проблемы – измерение характеристик почвы и непрерывное измерение урожайности. Например, в США, Германии, Англии, Дании и Франции налажен серийный выпуск аппаратуры для исследования вариабельности свойств почв и растительного покрова в пределах поля.

Их использование позволяет проводить дифференциальный учёт урожая, автоматически отбирать пробы почвы, оценивать состояние посевов с помощью съёмки из космоса и с высоты птичьего полета (аэрометоды).

Одним из примеров такого оборудования могут служить мобильные приборы, измеряющие электропроводность почвы. При перемещении по полю они позволяют измерять, наносить на карту и анализировать пространственную изменчивость тех или иных характеристик почвы. Однако и здесь существует ряд трудностей, так как величина измеряемого сигнала зависит от нескольких физических параметров почвы: гранулометрического состава, плотности сложения и влажности. По этой причине приходится применять и другие методы для того, чтобы вычленить влияние каждого из этих параметров.

Чрезвычайно простым представляется маятниковый прибор для измерения величины надземной массы. Но и в этом случае необходима калибровка по каждой культуре и даже по каждому сорту. Специальной калибровки требуют также оптические, в т.ч. лазерные, приборы, применяемые для решения той же самой задачи. Таким образом, при кажущемся разнообразии уже выпускаемых устройств, пригодных для нужд точного земледелия, их функциональные качества требуют совершенствования, так как все они грешат одним и тем же недостатком – в результат измерения приходится вносить поправки ввиду зависимости показаний от группы механических и физических величин. Исходя из этого, возникает задача комплексирования измерений, то есть построения такой физической модели почвы, которая бы позволила выявить и рассчитать вклад каждого отдельного параметра в измеряемый сигнал, например, содержания физической глины и песка, плотности почвы и её влажности.

Для исследования вариабельности свойств растительного покрова и почвы используются также многочисленные дистанционные методы. В основном это измерения в оптическом, гиперспектральном и радиолокационном диапазонах. С помощью этих методов можно определить, в частности, изменчивость почвенного покрова по содержанию

физической глины и органического вещества. Как бы ни были несовершенны эти методы в настоящее время, за ними будущее [2, 3, 7, 9].

Для обработки информации, получаемой с помощью информационно-измерительных систем, используются стационарные и бортовые компьютеры. Стационарный компьютер с программным обеспечением в общем случае должен выполнять следующие основные функции:

- ведение атрибутивной и пространственной базы данных с использованием геоинформационных систем (ГИС);
- ведение базы декларативных и процедурных знаний;
- обработку знаний и данных и формирование программы реализации информационной технологии ТЗ.

Бортовой компьютер с программным обеспечением должен выполнять следующие основные функции:

- фиксацию координат агрегатов (мобильных комплексов) в любой момент времени путём приёма сигналов от ГСП и других датчиков в процессе движения и осуществление при необходимости навигации в заданную точку;
- автоматическое создание электронных карт обследованных полей с разбивкой их на элементарные участки заданных размеров;
- обеспечение накопления и первичной обработки данных полевых измерений с использованием ГИС-технологий и экспорт этой первичной информации в стационарный компьютер;
- формирование управляющих сигналов для дифференцированного выполнения тех или иных агротехнических операций и обеспечение соответствующего их контроля на основе выработанной стационарным компьютером программы реализации технологии.

2.3. Технические средства для реализации технологии в полевых условиях

Для перехода от технологий, базирующихся на усредненных показателях параметров плодородия поля и состояния посевов, к избирательному воздействию на систему «почва – растение» необходимо, чтобы рабочие органы обрабатывающих орудий и сельскохозяйственных машин управлялись бортовым компьютером. Автоматизированная

система управления рабочими органами должна быть хорошо отлаженной, так чтобы агрегат чутко реагировал на изменение таких показателей, как норма высева, доза внесения удобрений и химических мелиорантов, расход средств химической защиты растений. От точности и надёжности техники такого рода зависит успех реализации точного земледелия.

В этой области в мире также существуют определённые достижения. Многие страны приступили к выпуску специальной сельскохозяйственной техники для точного земледелия, причём она специфична и способна производить работы на всех этапах возделывания сельскохозяйственной культуры – от предпосевной обработки почвы до уборки урожая. Параллельно ведутся интенсивные исследования по совершенствованию имеющейся техники и разработке новых машин и оборудования, отвечающих современным тенденциям развития новой информационной технологии.

Например, в Германии ещё в 1994 г. был создан комбайн, обеспечивающий дифференциальный учёт урожая зерновых культур в процессе уборки, а в 1997 г. в развитых странах стали массово выпускаться машины для избирательного внесения агрохимикатов с управлением от ГСП с использованием электронных карт поля, геоинформационных систем и бортовых компьютеров.

Особый интерес из всего комплекса разработанной техники представляют, конечно, комбайны, способные при уборке учитывать урожай. Такие машины используются для уборки корне- и клубнеплодов, существуют также зерновые комбайны, которые при уборке автоматически создают электронную карту урожайности с точной привязкой к месту. Такая карта попросту бесценна, так как нет более объективного показателя неоднородности пахотного массива, чем карта урожайности. Её наличие позволяет предметно судить о степени варьирования почвенно-климатических условий в пределах сельхозугодья и принимать обоснованные технологические решения. При этом появляется возможность более дифференцированно подходить к технологии возделывания сельскохозяйственной культуры на данном участке.

Товаропроизводители, агрономы и исследователи подразделяют факторы, влияющие на урожайность, на природные и антропогенные. Полезность создания карт зависит от того, насколько правильно они будут проанализированы. Главная цель интерпретации

карт урожайности – учесть все влияющие на урожай факторы и за счёт этого увеличить продуктивность агроценоза. Для лучшей расшифровки электронных карт привлекается дополнительная информация, помогающая установить связи между характеристиками угодья и величиной урожая.

Составление таких карт стало общей практикой сельскохозяйственного производства, например, в США. Некоторые поля имеют такую информацию уже за несколько лет, и количество мониторов по учету урожаев там постоянно растет [3].

3. КАРТЫ УРОЖАЙНОСТИ – ВАЖНАЯ ЧАСТЬ СБОРА И АНАЛИЗА ДАННЫХ ДЛЯ ПРИНЯТИЯ РЕШЕНИЙ

Развитие современных технологий позволяет получать важнейшую информацию о посевных площадях, т.н. карты урожайности. Используя специальные датчики, установленные на уборочной технике, а также бортовые компьютеры и приёмники GPS после уборки обмолачиваемых культур мы можем получать пространственно-ориентированные карты урожайности. Получение подобных карт является несомненным прорывом в области земледелия, так как позволяет нам определять неоднородность главного из всех показателей – урожайности.

Полученные карты включаются в геоинформационную базу хозяйства и служат отправной точкой при планировании агрохимического обследования, так как позволяют с высокой точностью выявлять проблемные участки поля. Эта информация существенно снижает издержки по обследованию поля, так как позволяет целенаправленно определять наиболее важные места для обследования.

На базе Меньковской опытной станции получали карты урожайности с помощью зернового комбайна C1aas Dominator 130 ©, оснащённого датчиками урожайности, бортовым компьютером и системой GPS. На рисунке 3.1 изображен пример такой карты, полученной после уборки ячменя в 2005 г. По карте можно достаточно точно определить границы участков с низкой урожайностью. В дальнейшем обследование почвы в этих местах целесообразно проводить особенно подробно, при этом важно выяснить причины низкой урожайности для того, чтобы в следующем сезоне избежать потерь урожая на этом поле.

Пространственное распределение урожайности имеет уникальное значение, так как нет более объективного показателя неоднородности сельскохозяйственного поля по плодородию, чем карта, характеризующая количественную интегральную оценку продукционного процесса. Подобная информация позволяет построить, в частности, различные алгоритмы выделения однородных участков (единиц управления) на заданном сельскохозяйственном поле. Рассмотрим разработанные подходы по оценке биоэквивалентности участков по урожайности.

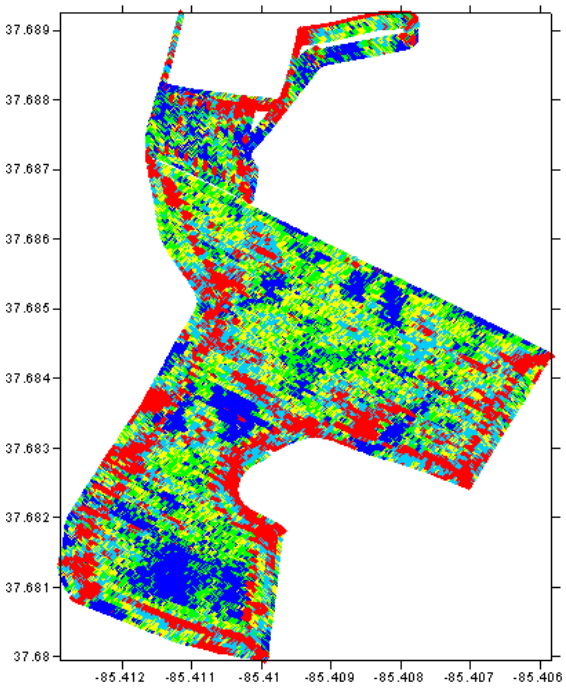

Рисунок 3.1. Карта урожайности

3.1. Алгоритмы выделения однородных участков по урожайности

Процесс получения растениеводческой продукции распределен во времени и пространстве на конкретной территории. Эта территория не является однородной даже в пределах одного поля или его части. По этой причине технологические операции, проводимые на данном поле, должны быть дифференцированы не только во времени (с учетом изменчивости погодных условий) и по полям севооборота, но и варьироваться в пределах одного поля.

Поле в этом случае может разделяться на некоторое количество новых единиц управления, которые являются однородными (квазиоднородными) участками. Однако выделение однородных участков на заданном поле представляет собой объективно сложную задачу. Рассмотрим далее два возможных подхода оценки степени однородности сельскохозяйственного поля по величине урожайности. Первый из них позволяет оценивать биоэквивалентность двух заданных участков на конкретном поле. Соответствующий алгоритм предполагает наличие данных по урожайности за ряд лет на сравниваемых участках. Второй подход, напротив, позволяет построить алгоритм выделения однородных зон на поле по величине урожайности его отдельных участков, полученной в конкретном году. Такая информация, как указывалось выше, фиксируется автоматически при уборке сельскохозяйственных полей с помощью комбайнов, оборудованных мони-

торами и соответствующими датчиками, с привязкой урожайности к глобальной системе координат.

3.1.1. Алгоритм оценки биоэквивалентности двух участков на сельскохозяйственном поле

Рассматриваемый алгоритм оценки биоэквивалентности двух участков по урожайности использует идею статистического моделирования выборок из нормальных распределений с заданными параметрами математических ожиданий и дисперсий по каждому из анализируемых участков. В биометрической литературе имеется большое количество работ, посвященных изучению биоэквивалентности. В некоторых работах было сформулировано понятие популяционной биоэквивалентности на основе анализа соответствующих экспериментальных данных.

Предположим, что на некотором поле можно выделить два участка A и B. Требуется принять решение о степени однородности или неоднородности этих участков по уровню средней урожайности на них (то есть о биологической эквивалентности). Обозначим средние урожайности на выделенных участках как Y_A (на участке A) и Y_B (на участке B). Предполагается, что каждый из участков состоит из большого числа небольших по площади делянок, а средняя урожайность представляет собой суммарный урожай всех делянок, поделенный на их количество. Урожайности на делянках различаются между собой и являются реализациями случайной величины, распределение которой неизвестно. Средняя урожайность участка также является случайной величиной, подчиняющейся некоторому распределению. Биологическая эквивалентность этих участков означает совпадение распределений этих случайных величин (средних урожайностей) или достаточную близость (сходство) этих распределений. В дальнейшем будем считать эти случайные величины взаимно независимыми. Подобное допущение представляется оправданным, если участки достаточно велики по площади.

Предположим, что имеются две выборки, представляющие собой значения средних урожайностей участков A и B за несколько предыдущих лет: $Y_{A1},...,Y_{An},Y_{B1},...,Y_{Bn}$. Представляется допустимым предположить, что распределения введенных случайных величин Y_A и Y_B являются нормальными с параметрами μ_A,μ_B (математические ожидания), $\delta^2{}_A,\delta^2{}_B$ (дисперсии), так как средние значения представляют собой нормированные

суммы урожайностей на небольших делянках, составляющих рассматриваемые участки, и, следовательно, при выполнении дополнительных предположений технического характера применима центральная предельная теорема Ляпунова.

Будем считать, что рассматриваемые участки биологически эквивалентны тогда, когда распределения случайных величин Y_A и Y_B, одинаковы или достаточно близки. Даже при одинаковых распределениях значения случайных величин, представляющих собой среднюю урожайность участков, в каждом конкретном испытании окажутся различными в силу предполагаемой независимости. В качестве меры сходства случайных величин Y_A и Y_B естественно выбрать вероятность

$$F_{A,B}(t) = P\{|\, Y_A - Y_B \,| \le r\} \qquad (1)$$

Если участки A и B биологически эквивалентны, то, как уже отмечалось, распределения средних урожайностей должны быть одинаковыми или достаточно близкими. Поэтому вероятность (1) следует сравнивать с вероятностью

$$F_{A,B}(t) = P\{|\, Y_A - Y'_A \,| \le r\} \qquad (2)$$

где случайная величина Y'_A независима от Y_A и имеет такое же распределение. В качестве величины r в формулах (1) и (2) возьмем $t\delta_A\sqrt{2}$. Тогда вероятности (1) и (2) можно переписать, используя функции стандартного нормального распределения, следующим образом:

$$F_{A,B}(t) = P\left\langle |\, Y_A - Y_B \,| \le t\sigma_A\sqrt{2} \right\rangle = \hat{O}\left\langle \frac{t\sigma_A\sqrt{2} + \mu_B - \mu_A}{\sqrt{\sigma^2 A + \sigma^2 B}} \right\rangle - \hat{O}\left\langle \frac{-t\sigma_A\sqrt{2} + \mu_B - \mu_A}{\sqrt{\sigma^2 A + \sigma^2 B}} \right\rangle$$

$$F_A(t) = P\left\langle |Y_A - Y_A| \le t\sigma_A\sqrt{2} \right\rangle = \hat{O}(t) - \hat{O}(-t)$$

где $\hat{O}(t)$ – функция стандартного нормального распределения.

Близость функций $F_{A,B}(t)$ и $F_A(t)$ свидетельствует о том, что участки A и B можно считать биоэквивалентными. Более строго биоэквивалентность можно определить в терминах разности или отношения введенных функций, для чего необходимо ввести пороговые значения f_d для разности и f_r для отношения.

Будем говорить, что участки A и B биоэквивалентны, если

$$F_{A,B}(t) - F_A(t) \ge f_d \text{ или } F_{A,B}(t) / F_A(t) \ge f_r \qquad (3)$$

Очевидно, что большее значение вероятности означает более высокую степень «близости» между средними урожайностями, что говорит о «более высокой биологической эквивалентности участков».

Используя имеющиеся выборки, построим оценки неизвестных параметров $\mu'_A, \mu'_B, \sigma'^2{}_A, \sigma'^2{}_B$, после чего, подставив их в функцию $F_{A,B}(t)$, получим статистическую оценку искомой вероятности: $F'_{A,B}(t)$. Далее можно производить сравнение найденной оценки с $F_A(t)$. Заметим, что функция $F_A(t)$ полностью известна. Проблема заключается в том, что вместо истинной вероятности $F_{A,B}(t)$ в сравнении используется статистическая оценка $F'_{A,B}(t)$.В связи с этим, можно предложить следующий алгоритм, использующий идею статистического моделирования выборок из нормальных распределений с параметрами $\mu'_A, \mu'_B, \sigma'^2{}_A, \sigma'^2{}_B$ (отдельно для участка A и участка B).

Алгоритм:

1) По полученным в результате многолетних наблюдений опытным данным строятся $\mu'_A, \mu'_B, \sigma'^2{}_A, \sigma'^2{}_B$.

2) На компьютере моделируются две выборки: выборка $Y_{A1},..., Y_{An}$ из нормального распределения с параметрами $\mu'_A, \delta'^2{}_A$ и выборка $Y_{B1},..., Y_{Bn}$ из нормального распределения с параметрами $\mu'_B, \delta'^2{}_B$.

3) По полученным выборкам строится новая оценка $F''_{A,B}(t)$, после чего для ранее выбранного порогового значения f_d (или f_r) производится проверка условия биоэквивалентности (3) (в форме разности или отношения).

4) Пункты 2 и 3 многократно повторяются, и подсчитывается относительная частота выполнения условия биоэквивалентности (3).

5) Принимается решение о принятии гипотезы биоэквивапентности участков или об отклонении этой гипотезы.

Рассмотрим применение предложенного алгоритма на примерах.

Пример 1. В примере моделируется разность средней урожайности каждого из участков за вычетом многолетней средней урожайности всего поля. Положим: $\mu_A = 0, \delta^2{}_A = 1, \mu_B = 0.5, \delta^2{}_B = 1$. Для смоделированной выборки опытных данных за 10 лет

(n=10) найдены оценки $\mu'_A = -0.49, \delta'^2_A = 0.46, \mu'_B = 0.19, \delta^2_B = 0.93$. В качестве порогового значения для разности вероятностей принято значение $f_d = -0.3$. Проведено 50 испытаний (пункты 2 и 3 алгоритма), в которых 40 раз выполнилось условие биоэквивалентности и 10 раз оно было нарушено. Ниже приведен вид графика разности функций вероятностей для одного из испытаний (рисунок 3.1.1.1), когда условие биоэквивалентности выполнено. Выборки содержали по 10 наблюдений. Результаты примера показывают, что, как правило, различия между участками A и B незначительны и, следовательно, они могут рассматриваться как биоэквивалентные.

Все вычисления производились в программе Maple 7:

$yA := stats[random, normald](10);$ (опытные данные для участка А)

$yB := stats[random, normald[0.5,1]](10);$ (опытные данные для участка В)

$f := (t) -> -0.3;$ (пороговое значение для разности вероятностей)

$plot[g - z, f], 0..8, color = [blue, red], style = [line, point];$

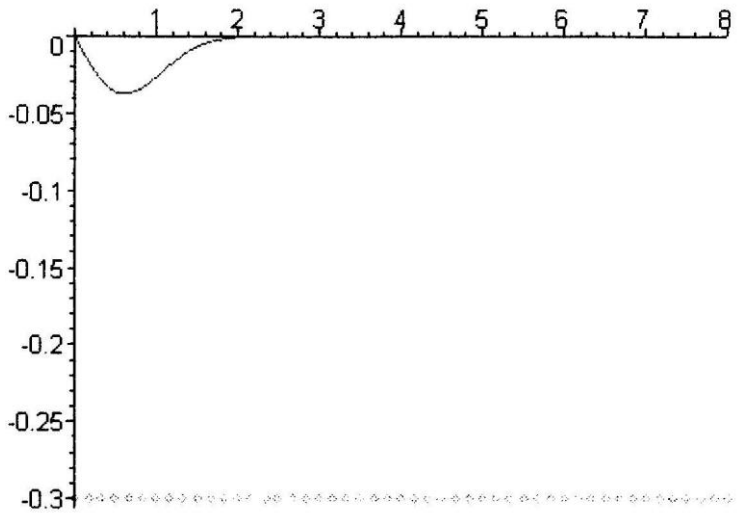

Рисунок 3.1.1.1. График разностей вероятностей с линией порогового значения для Примера 1.

Пример 2. В примере моделируется разность средней урожайности каждого из участков за вычетом многолетней средней урожайности всего поля. Положим: $\mu_A = 0, \delta^2_A = 1, \mu_B = 2.5, \delta^2_B = 1$. Для смоделированной выборки опытных данных за 10 лет $(n = 10)$ найдены оценки: $\mu'_A = -0.54, \delta^2_A = 0.84, \mu_B = 2.65, \delta^2_B = 0.38$. В качестве порогового значения для разности вероятностей принято значение $f_d = -0.3$. Проведено 50 испытаний (пункты 2 и 3 алгоритма), в которых 50 раз не выполнилось условие биоэквива-

лентности. Ниже приведен вид графика разности функций вероятностей для одного из испытаний (рисунок 3.1.1.2).Выборки содержали по 10 наблюдений.

Результаты примера показывают, что различия между участками A и B весьма значительны.

Все вычисления производились в программе Maple 7.

$yA := stats[random, normald](10);$ (опытные данные для участка А)

$yB := stats[random, normald[2.5,1]](10);$ (опытные данные для участка В)

$f := (t) - > -0.3;$ (пороговое значение для разности вероятностей

$plot[g - z, f], 0..8, color = [blue, red], style = [line, point];$

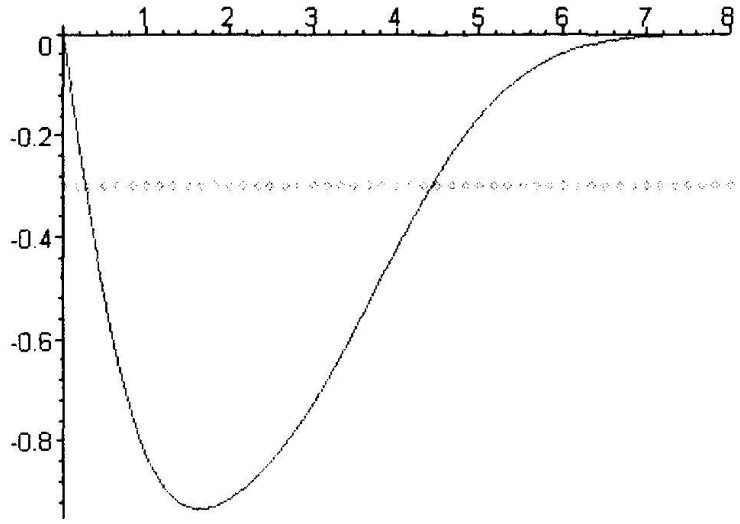

Рисунок 3.1.1.2. График разности вероятностей с линией порогового значения для Примера 2.

Рассмотренные примеры показывают работоспособность предлагаемого алгоритма. Алгоритм может быть реализован и в других программных продуктах, включая Excel, и позволяет при задаваемых различных уровнях урожайности проводить компьютерный анализ и оценку степени однородности тех или иных участков на заданном сельскохозяйственном поле, сравниваемых попарно.

3.1.2. Алгоритм выделения однородных зон на сельскохозяйственном поле по урожайности отдельных участков

Современные технологии ТЗ позволяют получать данные по урожайности с точной привязкой к координатам каждого отдельного участка на поле. Рассмотрим поле, состоящее из большого числа участков (размер участка – несколько квадратных метров), по каждому из которых измерена урожайность. На основе данных об урожайности по отдельным участкам для конкретного года требуется произвести разбиение поля на относи-

тельно однородные кластеры (зоны). Перенумеровав участки и зафиксировав урожайность на каждом отдельном участке, получаем массив данных, в котором содержится потенциальная информация об однородных зонах на поле. Предполагается, что урожайность внутри каждой из зон примерно одинакова, причем полное совпадение урожайностей, конечно, невозможно. Кроме того, при проведении нового опыта полной повторяемости урожайностей не будет, ввиду наличия многих факторов, оказывающих значимое воздействие на урожайность. Учитывая принципиальную неоднородность участков по свойствам почвы (отсутствие тождественности), возможное присутствие скрытых факторов, оказывающих влияние на урожайность отдельных участков, необходимо признать, что наличие близких значений урожайностей на двух участках в проведенном опыте не означает, что и в последующих опытах урожайности они будут такими же близкими. Скорее следует согласиться с тем, что они могут заметно различаться. Весь вопрос в том, как сильно будут они различаться, и как часто различия в урожайностях будут значительными. Урожайность на отдельных участках может моделироваться случайными величинами, подчиняющимися некоторому вероятностному распределению. Выбор конкретного распределения не очевиден, однако, применяя усреднение по количеству участков, можно воспользоваться предельными теоремами теории вероятностей [1].

Учитывая, что суммарная урожайность внутри зоны складывается из урожайностей большого количества отдельных участков, представляется допустимым считать, что средняя урожайность на одном участке внутри однородной зоны подчиняется нормальному распределению. Это предположение является ключевым для последующего анализа. Урожайности на соседних участках вряд ли могут считаться статистически независимыми, поэтому рассмотрение всего массива измерений для проведения статистического анализа вряд ли целесообразно.

Более целесообразным было бы учитывать урожайности отдельных участков, расположенных на расстоянии друг от друга таким образом, чтобы охватить все поле и обеспечить при этом относительную независимость урожайностей на них. Конечно, желательно, чтобы расстояние между участками было не слишком большим, поскольку по результатам измерений урожайности на этих участках будет осуществлена кластеризация поля. Ввиду небольших размеров участков их количество окажется значительным, что

позволяет обосновывать использование нормального распределения для описания закона распределения средней урожайности ссылкой на центральную предельную теорему Ляпунова [1], тем более что теперь отдельные слагаемые статистически независимы. Таким образом, рассмотрение средней урожайности на одном участке внутри однородной зоны в качестве генеральной совокупности позволяет обоснованно использовать нормальное распределение.

Конечно, имеющаяся выборка состоит из отдельных урожайностей на выбранных участках, поэтому ссылка на предельную теорему Ляпунова, вообще говоря, недостаточна. Однако обсуждаемая случайная изменчивость урожайности вызвана лишь пространственной неоднородностью участков. Если рассматривать участки внутри однородной зоны, то представляется естественным предполагать унимодальность и симметричность распределения урожайности на них, при этом внутри однородной зоны дисперсия не может быть большой. Все сказанное говорит о том, что распределение урожайности внутри однородной зоны близко к нормальному распределению.

Предположим, например, что все поле по величине урожайности можно условно разделить на пять кластеров (пять зон однородности). Первый кластер соответствует очень благоприятным условиям для произрастания данной культуры; второй – включает участки с хорошими условиями произрастания данной культуры; третий – соответствует в целом удовлетворительным условиям; четвертый – включает неблагоприятные зоны произрастания культуры и, наконец, пятый кластер включает участки с очень плохими условиями. Конечно, в действительности для конкретного поля количество однородных зон может быть меньше пяти, в идеальном случае все поле целиком может представлять собой одну однородную зону.

Кластеризация является одной из наиболее важных задач обработки данных. В настоящее время разработано большое количество методов и алгоритмов кластеризации, но, к сожалению, не все они могут эффективно работать с большими массивами данных, поэтому дальнейшие исследования в этом направлении связаны с преодолением этой проблемы. Одним из широко известных в аналитическом сообществе алгоритмов кластеризации, позволяющих эффективно работать с большими объемами данных, является EM-алгоритм.

Его название происходит от слов "expectation-maximization", что переводится как "ожидание-максимизация". Это связано с тем, что каждая итерация содержит два шага – вычисление математических ожиданий (expectation) и максимизацию (maximisation). Алгоритм основан на методике итеративного вычисления оценок максимального правдоподобия, предложенной в 1977 г.

В основе идеи EM-алгоритма лежит предположение, что исследуемое множество данных может быть смоделировано с помощью линейной комбинации многомерных нормальных распределений, а целью является оценка параметров распределения, которые максимизируют логарифмическую функцию правдоподобия, используемую в качестве меры качества модели. Иными словами, предполагается, что данные в каждом кластере подчиняются определенному закону распределения, а именно, нормальному распределению. С учетом этого предположения можно определить параметры – математическое ожидание и дисперсию, которые соответствуют закону распределения элементов в кластере, наилучшим образом "подходящему" к наблюдаемым данным.

Таким образом, мы предполагаем, что любое наблюдение принадлежит ко всем кластерам, но с разной вероятностью. Тогда задача будет заключаться в "подгонке" распределений смеси к данным, а затем в определении вероятностей принадлежности наблюдения к каждому кластеру. Очевидно, что наблюдение должно быть отнесено к тому кластеру, для которого данная вероятность выше.

Среди преимуществ EM-алгоритма можно выделить следующие:

- Мощная статистическая основа.
- Линейное увеличение сложности при росте объема данных.
- Устойчивость к шумам и пропускам в данных.
- Возможность построения желаемого числа кластеров.
- Быстрая сходимость при удачной инициализации.

Однако алгоритм имеет и ряд недостатков. Во-первых, предположение о нормальности всех измерений данных не всегда выполняется. Во-вторых, при неудачной инициализации сходимость алгоритма может оказаться медленной.

Кроме этого, алгоритм может остановиться в локальном минимуме и дать квазиоптимальное решение.

Статистические основы алгоритма

Как отмечалось во введении, EM-алгоритм предполагает, что кластеризуемые данные подчиняются линейной комбинации (смеси) нормальных (гауссовых) распределений. Плотность вероятности нормального распределения имеет вид:

$$p(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} \times \exp\{-\frac{(x-\mu)^2}{2 \times \sigma^2}\},$$

где $\mu = E(X)$ – математическое ожидание, $\sigma^2 = E(X-\mu)^2$ – дисперсия.

Многомерное нормальное распределение для q-мерного пространства является обобщением предыдущего выражения. Многомерная нормальная плотность для q-мерного вектора $x = (x_1, x_2, ..., x_q)$ может быть записана в виде:

$$p(x) = \frac{1}{(2 \times \pi)^{\frac{q}{2}} \times \sqrt{|\Sigma|}} \exp\{-\frac{1}{2} \times (x-\mu)^T \times \Sigma^{-1} \times (x-\mu)\},$$

где Σ – ковариационная матрица размером $q \times q$, которая, как известно, является обобщением дисперсии для многомерной случайной величины, μ представляет из себя q-мерный вектор математических ожиданий, $|\Sigma|$ – определитель ковариационной матрицы, T – оператор транспонирования.

Введем в рассмотрение функцию $\delta^2 = (x-\mu)^T \times \Sigma^{-1} \times (x-\mu)$, называемую квадратичным расстоянием Махаланобиса.

Алгоритм предполагает, что данные подчиняются смеси многомерных нормальных распределений для q переменных. Модель, представляющая собой смесь гауссовых распределений задается в виде:

$$p(x) = \sum_{i=1}^{k} w_i \times p(x \mid i),$$

где $p(x \mid i)$ – нормальное распределение для i-го кластера, w_i – доля (вес) i-го кластера в исходной базе данных.

Предположим, что каждый из k кластеров имеет свой вектор математических ожиданий μ, но все они имеют одну и ту же ковариационную матрицу Σ.

Существуют два подхода к решению задач кластеризации: основанный на расстоянии и основанный на плотности. Первый подход заключается в определении областей пространства признаков, внутри которых точки данных расположены ближе друг к другу, чем к точкам других областей, относительно некоторой функции расстояния (например, евклидовой). Второй – обнаруживает области, которые являются более "заселенными", чем другие. Алгоритмы кластеризации могут работать сверху вниз (иерархические) и снизу вверх (агломеративные). Агломаративные алгоритмы, как правило, являются более точными, хотя и работают медленнее.

Алгоритм EM основан на вычислении расстояний. Он может рассматриваться как обобщение кластеризации на основе анализа смеси вероятностных распределений. В процессе работы алгоритма происходит итеративное улучшение решения, а остановка осуществляется в момент, когда достигается требуемый уровень точности модели. Мерой в данном случае является монотонно увеличивающаяся статистическая величина, называемая логарифмическим правдоподобием. Целью алгоритма является оценка средних значений C, ковариаций R и весов смеси W для функции распределения вероятности, описанной выше. Параметры, оцененные алгоритмом, сохраняются в таблице вида:

Матрица	Размер	Содержит
C	$q \times k$	Математические ожидания
R	$q \times q$	Ковариации
W	$k \times 1$	Веса

Следует отметить, что один из популярных алгоритмов кластеризации k-means является частным случаем алгоритма EM, когда W и R постоянны: $w_i = \frac{1}{k}$, $R = I$ (I – единичная матрица).

Алгоритм начинает работу с инициализации, т.е. некоторого приближенного решения, которое может быть выбрано случайно или задано пользователем исходя из некоторых априорных сведений об исходных данных. Наиболее общим способом инициализации является присвоение элементам матрицы ма-

тематических ожиданий случайных значений ($C \leftarrow \mu Random$), начальная ковариационная матрица определяется как единичная ($R \leftarrow I$), веса кластеров задаются одинаковыми ($w_i \leftarrow \frac{1}{k}$).

Следует обратить внимание, что алгоритм может "застрять" в локальном оптимуме и дать квазиоптимальное решение при выборе неудачного начального приближения. Поэтому одним из его недостатков следует считать чувствительность к выбору начального состояния модели.

Реализация алгоритма EM может быть проиллюстрирована с помощью следующего псевдокода:

Вход: k – число кластеров,

$Y = \{y_1, y_2, ..., y_n\}$ – множество из наблюдений q-мерного пространства,

ε – допустимое отклонение для логарифмического правдоподобия,

Q – максимальное число итераций,

Выход: C, R, W – матрицы, содержащие обновляемые параметры смеси.

X – матрица с вероятностями членства в кластерах.

1. **Инициализация**: установка начальных значений C, R, W выбранных случайно или заданных пользователем.

2. **ПОКА** изменение логарифмического правдоподобия $\Delta llh \geq \varepsilon$ и не достигнуто максимальное число итераций Q, выполнять шаги **E** и **M**.

Шаг E

$C' = 0, R' = 0, W' = 0, llh = 0$

Для i, изменяющегося от 1 до n

$sump_i = 0$

Для j, изменяющегося от 1 до k

$$\delta_{ij} = (y_i - C_j)^T \times R^{-1} \times (y_i - C_j)$$

$$p_{ij} = \frac{w_j}{(2 \times \pi)^{\frac{q}{2}} \times |R|^{\frac{1}{2}}} \times \exp\{-\frac{1}{2} \times \delta_{ij}\}$$

$$sump_i = sump_i + p_{ij}$$

Конец цикла по j

$x_i = \frac{p_i}{sump_i}$, $llh = llh + \ln(sump_i)$

$C' = C' + y_i x_i^T$, $W' = W' + x_i$

Конец цикла по i

Шаг М

Для j, изменяющегося от 1 до k

$Cj = \frac{C'_j}{W'_j}$

Для i, изменяющегося от 1 до n

$R' = R' + (y_i - C_j) x_{ij} (y_i - C_j)^T$

Конец цикла по i

Конец цикла по j

$R = \frac{R'}{n}$, $W = \frac{W'}{n}$

Алгоритм содержит два шага: шаг ожидания (expectation) или **E**-шаг и шаг максимизации (maximization) или **M**-шаг. Каждый из них повторяется до тех пор, пока изменение логарифмического правдоподобия Δllh не станет меньше, чем ε, или пока не будет достигнуто максимальное число итераций.

Логарифмическое правдоподобие вычисляется как:

$$llh = \sum_{i=1}^{n} \ln(sump_i)$$

Переменные δ, R, P представляют собой матрицы, хранящие расстояния Махаланобиса, ковариации и вероятности членства в кластере для каждой из n точек. C', R' и W' являются временными матрицами, используемыми только для вычислений. $\| W \| = 1$, т.е. $\sum_{i=1}^{k} w_i$. Обозначение вида p_i, использованное в псевдокоде, обозначает k–размерный вектор принадлежности i-го наблюдения к каждому из k кластеров. Соответственно, x_i – нормированная вероятность принадлежности к каждому из k кластеров. Столбец C_j матрицы C матрицы есть оценка математического ожидания по j-му кластеру, R – диагональная

матрица, т.е. $R_{ij} = 0$ для всех $i \neq j$. Со статистической точки зрения это означает, что ковариации являются независимыми.

Диагональность является ключевым предположением, которое делает алгоритм масштабируемым. В этом случае детерминант матрицы и его обращение может быть вычислено за время $O(p)$, а алгоритм имеет сложность $O(kpn)$. В случае недиагональной матрицы сложность алгоритма составит $O(kp^2n)$, т.е. будет квадратично возрастать с увеличением размерности данных.

Важнейшим действием, выполняемым на **E**-шаге, является вычисление расстояний Махаланобиса δ_{ij}. Если матрица R является диагональной, то расстояние Махаланобиса от точки y до среднего значения кластера C, имеющего ковариацию R, будет:

$$\delta_{ij} = (y - C)^T R^{-1} (y - C) = \sum_{k=1}^{q} \frac{(y_k - C_k)^2}{C_{kk}}$$

поскольку $R_{kk}^{-1} = \frac{1}{R_{kk}}$, $\overline{k = 1, q}$. Если матрица R является диагональной, то ее обращение R^{-1} легко вычисляется, т.к. $(y_k - C_k) R_{kl}^{-1} = 0$ для любых $k \neq l$. Кроме этого, ускорению вычислений способствует то, что диагональная матрица R может храниться в виде вектора ее диагональных элементов. Поскольку R не изменяется в процессе **E**-шага, ее детерминант вычисляется только единожды, что делает вычисление вероятностей p_{ij} более быстрым. На **M**-шаге диагональность матрицы R также упрощает вычисления, поскольку недиагональные элементы матрицы $(y_i - C_j) x_{ij} (y_i - C_j)^T$ равны нулю.

Для оптимизации используемого объема памяти, алгоритм может работать в двух режимах. В первом загружается только часть доступных данных и на их основе предпринимается попытка построения модели. Если она увенчалась успехом, то алгоритм завершает работу, в противном случае загружается следующая порция данных и т.д., пока не будут получены приемлемые результаты. Во втором режиме загружаются сразу все имеющиеся данные. Как правило, последний вариант обеспечивает более точную подгонку модели, но предъявляет более жесткие требования к объему доступной оперативной памяти.

Численный эксперимент

Для иллюстрации работы алгоритма EM и его сравнения с k-means рассмотрим результаты численного эксперимента, для проведения которого была взята выборка, представленная на рисунке 3.1.2.1.

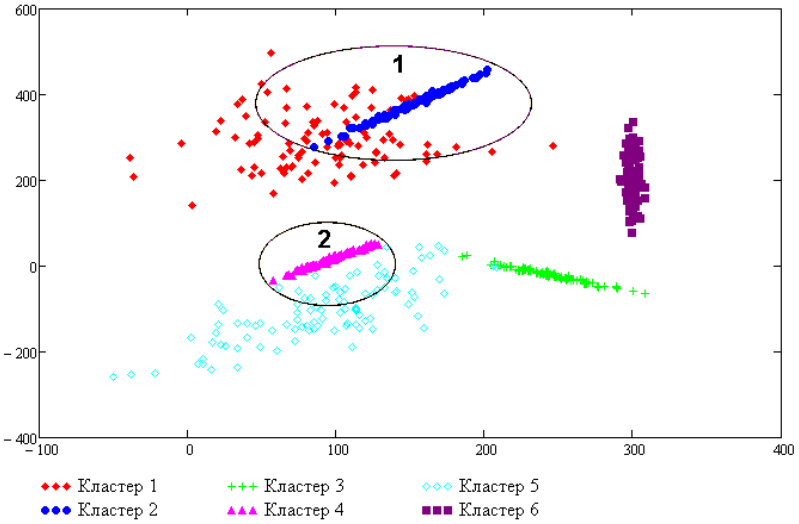

Рисунок 3.1.2.1. Исходные кластеры

Обратите внимание, что исходный набор данных, вообще говоря, не является простым с точки зрения задачи кластеризации, поскольку имеется явное перекрытие кластеров (области 1 и 2). В области 1 перекрываются кластеры 1 и 2, а в области 2 кластеры 4 и 5. Кластеры 3 и 6 расположены обособленно и, как ожидается, будут легко разделимы.

Для алгоритма k-means особые трудности должны возникнуть в местах перекрытия кластеров. Данное предположение подтверждается результатами, представленным на рисунке 3.1.2.2.

В местах перекрытия кластеров наблюдается наибольшее число ошибок. В то же время обособленные кластеры 3 и 6 были распознаны алгоритмом k-means без ошибок. Как можно увидеть на рисунке 3.1.2.3, алгоритм EM весьма успешно выявил перекрывающиеся кластеры, хотя и почти не распознал кластер 6.

Таким образом, можно сделать вывод, что алгоритм k-means может иметь преимущество при работе с обособленными (неперекрывающимися) кластерами, но полностью проигрывает алгоритму EM при наличии их перекрытия.

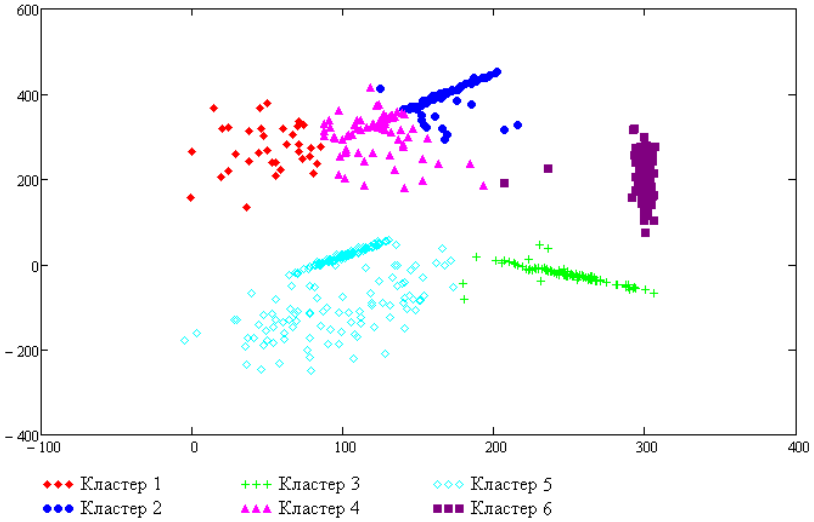

Рисунок 3.1.2.2. Результаты кластеризации k-means

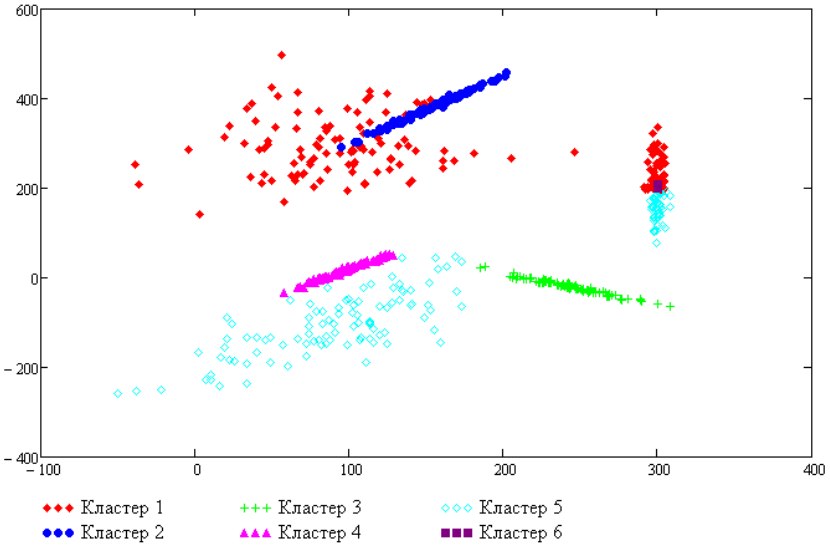

Рисунок 3.1.2.3. Результаты кластеризации EM

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы получены следующие результаты:

1. На основе биометрического подхода к определению биоэквивалентности двух относительно больших участков на сельскохозяйственном поле по урожайности некоторой сельскохозяйственной культуры за несколько лет с использованием статистического моделирования выборок из нормальных распределений разработан алгоритм оценки, выполнена его программная реализация.

2. Предложен метод выделения относительно однородных зон на сельскохозяйственном поле по урожайности отдельных небольших участков на поле за один год и дан вероятностный алгоритм решения этой задачи на основе разделения конечной смеси распределений. Выявлено, что использование тандема нескольких алгоритмов дает лучшие результаты.

Также хочется еще добавить:

1. Точное земледелие означает новый уровень учётной дисциплины в растениеводстве, базирующийся на компьютерных технологиях, цифровой картографии, спутниковой навигации и технологиях связи;

2. Внедрение технологий точного земледелия не может произойти единовременно, поскольку требуется построение инфраструктуры, компьютерных сетей, навигационной составляющей, построение сетей связи, подбор и обучение специалистов;

3. Внедрение технологий точного земледелия требует значительного объема доработок, а зачастую и переработок сельскохозяйственной техники и сельскохозяйственных агрегатов, выполнение которой наиболее целесообразно проводить на первичном рынке, т.е. при непосредственном участии отечественных предприятий изготовителей.

4. Анализ современной ситуации в Республике в области развития геоинформационных систем и спутниковой системы точного позиционирования, а также уровня научно-производственного потенциала показывает возможность максимального приближения к этой технологии в течение 2010-2015 годов

Из изложенного видно, что для внедрения технологий точного земледелия в Республике Беларусь необходим системный подход, предусматривающий комплексную разработку и внедрение всех элементов технологии точного земледелия. Эти работы могут быть выполнены в рамках, формируемых в настоящее время государственных комплексных целевых научно-технических программ (ГКЦНТП) на 2011-2015 г.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) Боровков А.А. Теория вероятностей. М.: Наука, 1986. 432 с.

2) Дринча В. М. Перспективные направления агроинженерных исследований для непрерывного устойчивого ведения сельского хозяйства. - М., ВИМ, 2004, 80 с.

3) Личман Г.И., Марченко Н.М., Дринча В.М. Основные принципы и перспективы применения точного земледелия. М., Россельхозакадемия, 2004, 80 с.

4) Семёнов В.А., Мирный В.И. Принципы адаптации технологий возделывания сельскохозяйственных культур. - В сб. «Программирование урожаев сельскохозяйственных культур на Северо-Западе РСФСР», Л., СЗНИИСХ, 1988, с. 4-9

5) Шатилов И.С., Чудновский А.Ф. Агрофизические, агрометеорологические и агротехнические основы программирования урожаев. Л.. Гидрометеоиздат, 1980, 320 с.

6) Якушев В.П. На пути к точному земледелию. - СПб.: Издательство ПИЯФ РАН, 2002, 458 с.

7) Якушев В.П., Полуэктов Р.А. и др. Точное земледелие: состояние исследований И задачи агрофизики. - В кн. «Агрофизические и экологические проблемы сельского хозяйства в 21 веке». СПб., SPBISTRO, 2002, т.3. С. 26-73

8) Якушев В.П., Полуэктов Р.А. Точное земледелие. Концептуальные положения. Материалы научной сессии Росссельхозакадемии (13-14 октября 2003 г.): «Научно- технический прогресс в АПК России - стратегия машинно-технологического обеспечения производства с/х продукции на период до 2010 г.». М., Россельхозакадемия, 2004, с. 115-123

9) Якушев В.П., Полуэктов Р.А. и др. Точное земледелие (аналитический обзор). Агрохимический вестник: №5, 2001, с. 28-34; №1, 2002, с. 34-39; №2, 2002, с. 36-39: №3, 2002, с. 36-40.

ПРИЛОЖЕНИЕ. КОДЫ ПРОГРАММ ДЛЯ ПУНКТА 3.1.2. АЛГОРИТМ ВЫДЕЛЕНИЯ ОДНОРОДНЫХ ЗОН НА СЕЛЬСКОХОЗЯЙСТВЕННОМ ПОЛЕ ПО УРОЖАЙНОСТИ ОТДЕЛЬНЫХ УЧАСТКОВ

kMeansParam.m

```
function y=kMeansParam(X,k,isRand)
if nargin<3, isRand=0; end
if nargin<2, k=1; end
[maxRow, maxCol]=size(X);
if maxRow<=k,
y=[X, (1:maxRow)'];
else if isRand,
p=randperm(maxRow);
c=X(p(1:k),:);
else c=X((1:k),:);
end
temp=zeros(maxRow,1);
while 1,
d=DistMatrix(X,c);
[z,g]=min(d,[],2); % find group matrix g
if g==temp, break;
else temp=g; end
for i=1:k
f=find(g==i);
if f
c(i,:)=mean(X(find(g==i),:),1);
end
end
end
p=zeros(k,1); % pre-initialization for speed
for i=1:maxRow
p(g(i))=p(g(i))+1;
end
q.p=zeros(k,1);
q.p(:)=p(:)./maxRow;
```

```
q.m=c;
for i=1:k
h=find(g==i);
if h
c=X(find(g==i),:);
end
diffs = c - ones(size(c,1),1)*q.m(i,:);
q.e{i}=(diffs'*diffs)/size(c,1);
if rank(q.e{i})<maxCol
q.e{i} = q.e{i}+ 0.000001*eye(maxCol);
end
end
y=q;
end
```

ExpMin.m

```
function y=ExpMin(X,PP)
if (nargin < 2)
PP.dummy=0;
end
if ~isfield(PP, 'k')
PP.k=1;
end
if ~isfield(PP, 'toStop')
PP.toStop = 0.001;
end
k = PP.k;
toStop = PP.toStop;
[maxRow, maxCol]=size(X);
if maxRow<=k,
y=[X, (1:maxRow)'];
else
q=kMeansParam(X,k,1);
t=1;
g=zeros(maxRow,k);
while t>toStop,
```

```
t=0;
gg=g;
r=zeros(maxRow,k);
for l=1:maxRow
for j=1:k
r(l,j)=(q.p(j))*(mvnpdf(X(l,:),q.m(j,:),q.e{j}+ 0.000001*eye(maxCol)));
end
g(l,:)=r(l,:)./(sum(r(l,:)));
end
t=max(max(max(abs(g-gg),t)));
n=zeros(k,1);
for j=1:k
n(j)=sum(g(:,j));
end
q.p(:)=n(:)/maxRow;
for j=1:k
f=zeros(maxRow,maxCol);
for l=1:maxRow
f(l,:)=g(l,j)*X(l,:);
end
q.m(j,:)=(sum(f))/(n(j));
f=zeros(maxCol,maxCol);
for l=1:maxRow
f=f+((X(l,:)-q.m(j,:))'*(X(l,:)-q.m(j,:)))*g(l,j);
end
q.e{j}=f/(n(j));
end
end
q.x = X;
for l=1:maxRow
q.y(l) = find(g(l,:)== max(g(l,:)));
end
y=q;
end
```

DistMatrix.m

```
function d=DistMatrix(A,B)
[hA,wA]=size(A);
[hB,wB]=size(B);
if wA~= wB, error(' second dimension of A and B must be the same');
end
for k=1:wA
C{k}= repmat(A(:,k),1,hB);
D{k}= repmat(B(:,k),1,hA);
end
S=zeros(hA,hB);
for k=1:wA
S=S+(C{k}-D{k}').^2;
end
d=sqrt(S);
```

main.m

```
N=100;
alpha=2.5;
sig2=1;
dist2=1.5;
N2=floor(alpha * N);
cls1X=randn(N, 2);
ShiftClass2=repmat(dist2 * [sin(pi*rand) cos(pi*rand)],[N2,1]);
cls2X = sig2 * randn(N2, 2) + ShiftClass2;
X = [cls1X; cls2X];
y = [zeros(size(cls1X,1),1); ones(size(cls2X,1),1)];
X; y;
idx1 = find(y == 0);
idx2 = find(y == 1);
h = figure; hold on
plot(X(idx1,1), X(idx1,2), 'r*');
plot(X(idx2,1), X(idx2,2), 'b*');
axis tight
xlabel('x_1');
ylabel('x_2');
k=2;
```

```
w = ExpMin(X,k,2);
idx1 = find(w.y == 1);
idx2 = find(w.y == 2);
plot(w.x(idx1,1), w.x(idx1,2), 'ro');
plot(w.x(idx2,1), w.x(idx2,2), 'bo');
```