На правах рукописи

Векшина Анна Борисовна

РАЗРАБОТКА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АДАПТИВНОЙ МОДЕЛИ ГЕНОГЕОГРАФИЧЕСКОГО ПРОГНОЗА НА ОСНОВЕ МЕТОДОВ ОПТИМАЛЬНОГО ОЦЕНИВАНИЯ И ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

05.13.18 – Математическое моделирование,
численные методы и комплексы программ,

05.11.17 – Приборы, системы и изделия медицинского назначения

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технически наук

Москва-2012

Работа выполнена на кафедре 704 «Информационно-управляющие комплексы» Московского авиационного института (национального исследовательского университета, МАИ)

Научный руководитель: доктор технических наук, профессор
Евдокименков Вениамин Николаевич

Научный консультант: доктор медицинских наук, профессор
Зинченко Рена Абульфазовна

Официальные оппоненты: Падалко Сергей Николаевич, доктор технических наук, профессор, заместитель заведующего кафедрой 609 «Прикладная информатика» Московского авиационного института (национального исследовательского университета, МАИ)

Филист Сергей Алексеевич, доктор технических наук, профессор, заместитель заведующего кафедрой БМИ «Биомедицинской инженерии» Юго-Западного государственного университета (ЮЗГУ)

Ведущая организация: Федеральное государственное бюджетное учреждение науки Государственный научный центр Российской Федерации – Институт медико-биологических проблем Российской академии наук (ГНЦ РФ-ИМБП РАН)

Защита состоится «_25_»__мая__ 2012г. в _13.00_ часов на заседании диссертационного совета Д 212.125.12 в Московском авиационном институте (национальном исследовательском университете, МАИ) по адресу: 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4.

С диссертацией можно ознакомиться в библиотеке Московского авиационного института (национального исследовательского университета, МАИ) по адресу: 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4.

Автореферат разослан «_24_»_апреля_ 2012 г.

Отзывы, заверенные печатью, просьба высылать по адресу: 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4, МАИ, Учѐный совет МАИ.

Учѐный секретарь диссертационного совета Д 212.125.12,
кандидат технических наук, доцент В.В. Дарнопых

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Основными задачами современной медицинской генетики являются профилактические программы, включающие такие основные компоненты, как выявление суммарного груза и разнообразия наследственных болезней в популяциях, а также оценка основных механизмов их формирования и распространения. Решение этих задач невозможно без оценки выраженной географической изменчивости отдельных наследственных болезней по регионам и распределения значений генетических показателей в границах исследуемых популяций.

В настоящее время основным источником знаний о распространенности наследственных патологий и распределении значений генетических показателей в границах некоторой популяции являются экспедиционные популяционно-генетические исследования, в процессе которых формируются данные о спектре наследственных заболеваний, преобладающих в популяции и о суммарном генетическом грузе в целом. Однако, проведение такого рода исследований, предполагающих организацию экспедиций с привлечением специального оборудования и высококвалифицированных специалистов, осложняется объективно существующими временными и материальными ограничениями. Иными словами, данные о популяции, как правило, представлены результатами ограниченного объема клинико-биохимических и молекулярно-генетических исследований, проведенных в конкретных населенных пунктах. Такого рода информация не позволяет получить полное представление о географической изменчивости значений тех или иных генетических показателей в границах исследуемой популяции (Ю.Г. Рычков, 2002; Е.В. Балановская, О.В. Балановский, 2007).

Выходом из подобной ситуации является разработка и внедрение в практику популяционно-генетических исследований математических моделей, позволяющих на основе ограниченного объема данных, полученных в ходе фактически проведенных экспедиционных исследований прогнозировать значения интересующих специалистов генетических показателей в любом населенном пункте в границах популяции.

Существует ряд работ, посвященных исследованиям в данной предметной области, однако описанные в работах модели геногеографического прогноза ориентированы на прогнозирование значений генетических показателей человека, характеризующих условно нормальную часть генома (A.J. Ammerman, L.L. Cavalli-Sforza, 1984; Е.В. Балановская, С.Д. Нурбаев, Ю.Г. Рычков, 1994). Использование такого рода моделей для прогноза значений генетических показателей, связанных с наследственными заболеваниями, которые являются редкими событиями и характеризуют патологическую часть генома, не обеспечивает надлежащей точности результатов, из-за возникновения ошибок, обусловленных сложностью биологических процессов, приводящих к развитию наследственных патологий.

Учитывая вышеизложенное, актуальной задачей является создание математической модели для прогноза значений генетических показателей, позволяющей корректно формировать оценки, связанные с наследственными болезнями человека, в любых населенных пунктах исследуемой популяции. Данные, полученные с помощью такой модели, помогут выявить районы с высоким риском заболеваний, связанных с теми или иными генами, укажут области с рас-

пространением тех генетических свойств, которые имеют значение при переливании крови, при трансплантации органов и тканей.

Кроме того, экспериментальные и прогнозные, полученные с помощью математического моделирования, значения генетических показателей служат основой для построения геногеографических карт, которые с приемлемой степенью точности заполняют обширные пробелы в знаниях о генетике населения и служат источником предварительной генетической информации.

Объект исследования – популяционные закономерности распространения генетических заболеваний. **Предмет исследования** – геногеографические модели для оценки распределения значений генетических показателей.

Цель и задачи диссертационной работы. Основной целью диссертационной работы является повышение информативности и достоверности результатов популяционно-генетических исследований на основе разработки и внедрения математических моделей и реализующих их программных комплексов, которые обеспечивают получение значений генетических показателей в населенных пунктах исследуемой популяции, не охваченных экспедиционными популяционно-генетическими исследованиями.

Для достижения этой цели были поставлены и решены следующие **основные задачи**:

1) анализ и обобщение существующих подходов к разработке моделей геногеографического прогноза;

2) разработка адаптивной математической модели геногеографического прогноза на основе ограниченного количества фактически полученных данных, обеспечивающей получение оценок генетических показателей в любом населенном пункте исследуемой популяции на основании его географических координат и численности проживающего населения. Адаптивность разрабатываемой модели предполагает автоматическую реконфигурацию ее структуры и уточнение параметров модели в зависимости от объема результатов экспедиционных популяционно-генетических исследований, привлекаемых для настройки модели;

3) создание алгоритмов расчета параметров модели геногеографического прогноза на основе данных экспедиционных популяционно-генетических исследований;

4) разработка метода формирования оптимального плана проведения экспедиционных популяционно-генетических исследований, использование которого гарантирует последующее получение модели геногеографического прогноза, обладающей максимальной точностью;

5) разработка программного комплекса, реализующего планирование экспедиционных популяционно-генетических исследований, прогноз значений генетических показателей, на основе адаптивной геногеографической модели и возможность визуализации полученных результатов на географической карте;

6) оценка эффективности разработанной адаптивной модели геногеографического прогноза и программного комплекса на примере обработки и анализа результатов экспедиционных популяционно-генетических исследований, охватывающих различные регионы России.

Методы исследования. При достижении целей исследования были использованы фундаментальные методы оптимального стохастического оцени-

вания, методы функционального анализа, методы теории планирования эксперимента, методы теории вероятности и математической статистики, а так же технология объектно-ориентированного программирования.

Основные положения диссертационной работы, выносимые на защиту:

1) адаптивная математическая модель геногеографического прогноза, позволяющая получать значения генетических показателей в рамках исследуемой популяции. Отличие разработанной модели от известных аналогов состоит в том, что ее структура и параметры автоматически настраиваются в зависимости от объема фактически проведенных популяционно-генетических исследований, доступных для анализа;

2) комплекс алгоритмов, обеспечивающих автоматическую адаптацию и расчет параметров моделей геногеографического прогноза на основе результатов популяционно-генетических исследований;

3) метод построения *D*-оптимального плана для выбора населенных пунктов, являющихся объектами экспедиционных популяционно-генетических исследований, который обеспечивает получение модели геногеографического прогноза обладающей максимальной точностью;

4) программный комплекс *GEN*, который обеспечивает получение оценок генетических показателей в границах исследуемых популяций на основе разработанных моделей геногеографического прогноза и их визуализация путем представления полученных результатов на географической карте.

Научная новизна. В процессе решения поставленных задач получены следующие новые научные результаты:

1) разработана адаптивная модель геногеографического прогноза, обеспечивающая двукратное повышение точности прогноза по сравнению с моделями неизменной линейной структуры и учитывающая в процессе прогноза не только географические координаты административно-территориальных образований, но и численность, проживающего в них населения, что дает возможность более точного прогноза значений генетических показателей человека, связанных с наследственными заболеваниями;

2) создан комплекс алгоритмов, позволяющих осуществлять автоматическую адаптацию структуры моделей геногеографического прогноза и расчет их параметров в зависимости от объема доступных для анализа результатов экспедиционных популяционно-генетических исследований. Преимущество разработанных алгоритмов заключается в автоматической настройке структуры и параметров модели прогноза, что позволяет исключить какое-либо субъективное влияние на достоверность результатов прогноза со стороны пользователей (специалистов-генетиков), не обладающих достаточной математической подготовкой;

3) разработан метод формирования *D*-оптимального плана экспедиционных популяционно-генетических исследований, позволяющий в условиях объективного наличия временных и материальных ресурсов наилучшим образом в смысле точности модели геногеографического прогноза выбрать населенные пункты для проведения экспедиционных популяционно-генетических исследований;

4) разработан и реализован в среде *Delphi* программный комплекс *GEN*, основу которого составляют разработанные модели геногеографического прогноза и оптимального планирования экспедиционных исследований с возможностью графического представления полученных результатов на географической карте.

Практическая значимость работы и результаты внедрения.

1) Созданный программный комплекс *GEN* обеспечивает на основе ограниченного объема фактически проведенных популяционно-генетических исследований получение прогнозных значений интересующего специалистов генетического показателя в тех населенных пунктах, где исследования не проводились;

2) Реализация в структуре комплекса *GEN* алгоритмов оптимального планирования позволяет в условиях временных и материальных ограничений обоснованно выбирать населенные пункты для проведения экспедиционных исследований, таким образом, чтобы построенная на их основе модель геногеографического прогноза обладала максимальной точностью.

3) Основные результаты диссертационной работы внедрены в Федеральном государственном бюджетном учреждении «Медико-генетический научный центр» Российской академии медицинских наук в процессе планирования популяционно-генетических исследований, обработки и анализа их результатов и в учебном процессе Московского авиационного института по специальности 200402 «инженерное дело в медико-биологической практике», что подтверждается соответствующими актами.

Достоверность результатов, полученных в диссертационной работе, подтверждается использованием аппарата математической статистики, оптимального планирования эксперимента; сопоставлением результатов, полученных с помощью разработанной математической модели, с данными экспедиционных исследований, охватывающих большое число популяций России (Ростовская область, Кировская область, Республика Чувашия, Республика Удмуртия, Республика Мари Эл и др.); значительным объемом выполненных в работе вычислений, результаты которых являются непротиворечивыми и укладываются в рамки существующих представлений теории оптимизации и планирования эксперимента.

Апробация работы. Основные положения диссертационной работы обсуждались и докладывались на 10-ой международной конференции «Системный анализ, управление и навигация» (Крым, Евпатория, 2005), 4-ой международной конференции «Авиация и космонавтика-2005» (Россия, Москва, 2005), Европейской конференции по генетике человека 2009 (Вена, 2009), 1-ой международной научно-практической конференции «Достижения, инновационные направления, перспективы развития и проблемы современной медицинской науки, генетики и биотехнологий» (Россия, Екатеринбург, 2011), 9-ой международной научно-практической конференции «Интеллект и наука» (Россия, Железногорск, 2011), на IV Всероссийской научно-практической конференции с международным участием «Биомедицинская инженерия и биотехнология» - г.Курск, КГМУ.

Публикации. Основные результаты диссертационной работы опубликованы в [1-3] журналах, входящих в рекомендованный ВАКом Минобрнауки Рос-

сии перечень изданий, одна работа [4] в зарубежном издании и пять работ [5-9] в сборниках тезисов докладов на научно-технических конференциях.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав основного материала, заключения и списка литературы из 108 наименований. Общий объем работы составляет 127 страниц основного текста, в том числе 51 рисунок и 26 таблиц.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность темы исследований, определена цель диссертационной работы и приведено ее краткое содержание.

В первой главе проведен обзор современного состояния исследований в области геногеографии и анализ существующих математических моделей прогноза значений генетических показателей. Рассмотрены основные методы картографирования и описаны принципы создания геногеографических карт.

Анализ сложившихся в настоящее время подходов к геногеографическому прогнозированию на основе математических моделей показал, что известные варианты реализованных математических моделей не обладают возможностью их автоматической адаптации с увеличением объема результатов популяционно-генетических исследований. Сегодня для прогноза значений генетических показателей применяются модели на основе линейных полиномов, структура которых не зависит от объема доступных для анализа результатов экспедиционных исследований. В литературе указывается возможность использования полиномов более высокой степени для решения специальных задач, требующих увеличения точности моделирования, однако отсутствуют методы их автоматической реконфигурации в процессе исследования (Ю.Г. Рычков, 2000, 2002). Выбор степени полинома при решении задач такого рода никак не регламентируется и возлагается на специалиста-генетика, что снижает достоверность результатов прогноза и делает их подверженными субъективному влиянию, обусловленному квалификацией специалиста-генетика.

Кроме того, ни одна из существующих геногеографических моделей не учитывает такого важного с точки зрения прогноза значений генетических показателей, связанных с распространенностью наследственных заболеваний, фактора, как численность населения в исследуемом регионе. Не учет численности населения может привести к ошибкам, поскольку при прогнозе модель будет основываться исключительно на данных о локализации исследуемых областей (географические долгота и широта), которые не несут информацию о миграционной активности населения, а, следовательно, и генетической неоднородности исследуемых областей. Модели, не учитывающие численность населения при прогнозе, успешно используется для получения значений генетических показателей человека в норме. Однако применять их для прогнозирования, связанного с наследственными болезнями, не корректно, поскольку в данном случае полиморфный уровень (частота гена) имеет другой порядок.

Анализ литературных источников показывает, что ни в одной из работ не описан подход по формированию набора населенных пунктов для проведения в них экспедиционных популяционно-генетических исследований, с целью построения на этих данных модели геногеографического прогноза. То есть специалисты сами на основе интуиции, историко-культурных и этнографических знаний о популяции выбирают населенные пункты для проведения популяци-

онно-генетических исследований. Таким образом, выбор связан с экспертным анализом специалистом-генетиком значимости ряда факторов для генофонда, что является субъективной оценкой и может привести к снижению точности при прогнозировании генетических показателей с помощью, полученной на этих данных, математической модели.

На основе проведенного анализа сформулированы основные задачи исследования.

Вторая глава диссертационной работы посвящена разработке адаптивной модели геногеографического прогноза и комплекса по оптимизации плана экспедиционных популяционно-генетических исследований.

Формализация модели прогноза значений генетических показателей опирается на доказанные геногеографические закономерности, предполагающие в границах исследуемой популяции функциональную связь между распространенностью заболевания (числом больных N^*), численностью населения N и географическим расположением административно-территориального образования (с координатами φ, λ), то есть, на существование зависимости вида:

$$N^* = N^*(N, \varphi, \lambda) \quad (1)$$

Задача разработки модели геногеографического прогноза формулируется следующим образом. Предполагается, что для анализа доступны результаты экспедиционных популяционно-генетических исследований, проведенных в ограниченном числе $i=1,..,m$ населенных пунктов популяции. Эти результаты включают в себя следующие данные: φ_i ,λ_i – соответственно широта и долгота населенного пункта, в котором проведены генетические исследования, N_i – численность населения, проживающего в этом административно-территориальном образовании, N_i^* – выявленное число носителей определенной наследственной патологии. Необходимо синтезировать зависимость (1), которая позволяет на основе известных данных по численности населения N в любом административно-территориальном образовании в границах исследуемой популяции и его географической локализации φ, λ оценить распространенность наследственного заболевания N^*.

Для получения модели геногеографического прогноза использовано разложение функциональной зависимости $N^*(N, \varphi, \lambda)$ в ряд Тейлора в окрестности одного из известных значений $N_i^*=N^*(N_i, \varphi_i, \lambda_i)$, $i=1,...,m$, полученных в ходе экспедиционных популяционно-генетических исследований i-го административно-территориального образования. В общем случае при наличии спектральных данных по m эталонным объектам возможно построение модели с переменными коэффициентами типа (1), за счет использования членов разложения в ряд Тейлора порядка $h = \left[\frac{m-1}{3}\right]$, позволяющих учесть производные порядков до h включительно, которые характеризуют изменение значений интересующего нас генетического показателя с учетом широты, долготы административно-территориальной единицы (село, город, район) и численности проживающего на ее территории населения. В вышеприведенном выражении $\left[\frac{m-1}{3}\right]$ - результат округления значения $\frac{m-1}{3}$ до целого в меньшую сторону. Подобная общая модель геногеографического прогноза приобретает вид:

$$N^*(N,\varphi,\lambda) = N^*(N_i,\varphi_i,\lambda_i) + (\frac{\partial N^*}{\partial N})_i(N-N_i) + (\frac{\partial N^*}{\partial \varphi})_i(\varphi-\varphi_i) + (\frac{\partial N^*}{\partial \lambda})_i(\lambda-\lambda_i) + (\frac{\partial^2 N^*}{\partial N^2})_i(N-N_i)^2 + \\ + (\frac{\partial^2 N^*}{\partial \varphi^2})_i(\varphi-\varphi_i)^2 + (\frac{\partial^2 N^*}{\partial \lambda^2})_i(\lambda-\lambda_i)^2 + \ldots + (\frac{\partial^h N^*}{\partial N^h})_i(N-N_i)^h + (\frac{\partial^h N^*}{\partial \varphi^h})_i(\varphi-\varphi_i)^h + (\frac{\partial^h N^*}{\partial \lambda^h})_i(\lambda-\lambda_i)^h \quad (2)$$

Для получения оптимальных оценок $3hm$ производных

$$(\frac{\partial N^*}{\partial N})_i, (\frac{\partial N^*}{\partial \varphi})_i, (\frac{\partial N^*}{\partial \lambda})_i, (\frac{\partial^2 N^*}{\partial N^2})_i, (\frac{\partial^2 N^*}{\partial \varphi^2})_i, (\frac{\partial^2 N^*}{\partial \lambda^2})_i, \ldots, (\frac{\partial^h N^*}{\partial N^h})_i, (\frac{\partial^h N^*}{\partial \varphi^h})_i, (\frac{\partial^h N^*}{\partial \lambda^h})_i, i = 1,\ldots,h,$$

в соответствии с методом наименьших квадратов (МНК) используется выражение

$$\hat{a} = (F^T F)^{-1} F^T Y, \quad (3)$$

в котором

$$a = \left((\frac{\partial N^*}{\partial N})_1, (\frac{\partial N^*}{\partial \varphi})_1, (\frac{\partial N^*}{\partial \lambda})_1, \ldots, (\frac{\partial^h N^*}{\partial N^h})_1, (\frac{\partial^h N^*}{\partial \varphi^h})_1, (\frac{\partial^h N^*}{\partial \lambda^h})_1 \ldots, (\frac{\partial N^*}{\partial N})_m, (\frac{\partial N^*}{\partial \varphi})_m, (\frac{\partial N^*}{\partial \lambda})_m, \ldots, (\frac{\partial^h N^*}{\partial N^h})_m, (\frac{\partial^h N^*}{\partial \varphi^h})_m, (\frac{\partial^h N^*}{\partial \lambda^h})_m \right)^T \quad (4)$$

вектор размерности $3hm$ x 1, компонентами которого являются оцениваемые производные.

$$Y = \begin{pmatrix} N^*(N_2,\varphi_2,\lambda_2) - N^*(N_1,\varphi_1,\lambda_1) \\ N^*(N_3,\varphi_3,\lambda_3) - N^*(N_1,\varphi_1,\lambda_1) \\ \text{----------} \\ N^*(N_m,\varphi_m,\lambda_m) - N^*(N_1,\varphi_1,\lambda_1) \\ N^*(N_1,\varphi_1,\lambda_1) - N^*(N_2,\varphi_2,\lambda_2) \\ N^*(N_3,\varphi_3,\lambda_3) - N^*(N_2,\varphi_2,\lambda_2) \\ \text{----------} \\ N^*(N_m,\varphi_m,\lambda_m) - N^*(N_2,\varphi_2,\lambda_2) \\ \text{----------} \\ N^*(N_{m-1},\varphi_{m-1},\lambda_{m-1}) - N^*(N_m,\varphi_m,\lambda_m) \end{pmatrix}$$ - вектор размерности $m(m-1) \times 1$, каждая

компонента которого представляет собой попарные комбинации разностей значений генетического показателя. Матрица F имеет размер $m(m-1) \times 3hm$ и следующее блочное представление:

$$F = \begin{pmatrix} F_{11} & F_{12} & \ldots & F_{1m} \\ F_{21} & F_{22} & \ldots & F_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ F_{m1} & F_{m2} & \ldots & F_{mm} \end{pmatrix} \quad (5)$$

Диагональные блоки F_{11}, F_{22} ,…, F_{mm} представляют собой матрицы размера $(m-1) \times 3h$ и имеют структуру:

$$F_{11} = \begin{pmatrix} (N_2-N_1) & (\varphi_2-\varphi_1) & (\lambda_2-\lambda_1) & (N_2-N_1)^2 & (\varphi_2-\varphi_1)^2 & (\lambda_2-\lambda_1)^2 & \ldots\ldots\ldots (N_2-\varphi_1)^h & (\varphi_2-\varphi_1)^h & (\lambda_2-\lambda_1)^h \\ (N_3-N_1) & (\varphi_3-\varphi_1) & (\lambda_3-\lambda_1) & (N_3-N_1)^2 & (\varphi_3-\varphi_1)^2 & (\lambda_3-\lambda_1)^2 & \ldots\ldots\ldots (N_3-N_1)^h & (\varphi_3-\varphi_1)^h & (\lambda_3-\lambda_1)^h \\ \text{----------} \\ (N_m-N_1) & (\varphi_m-\varphi_1) & (\lambda_m-\lambda_1) & (N_m-N_1)^2 & (\varphi_m-\varphi_1)^2 & (\lambda_m-\lambda_1)^2 & \ldots\ldots\ldots (N_m-N_1)^h & (\varphi_m-\varphi_1)^h & (\lambda_m-\lambda_1)^h \end{pmatrix}$$

$$F_{22} = \begin{pmatrix} (N_1-N_2) & (\varphi_1-\varphi_2) & (\lambda_1-\lambda_2) & (N_1-N_2)^2 & (\varphi_1-\varphi_2)^2 & (\lambda_1-\lambda_2)^2 & \ldots\ldots\ldots & (N_1-N_2)^h & (\varphi_1-\varphi_2)^h & (\lambda_1-\lambda_2)^h \\ (N_3-N_2) & (\varphi_3-\varphi_2) & (\lambda_3-\lambda_2) & (N_3-N_2)^2 & (\varphi_3-\varphi_2)^2 & (\lambda_3-\lambda_2)^2 & \ldots\ldots\ldots & (N_3-N_2)^h & (\varphi_3-\varphi_2)^h & (\lambda_3-\lambda_2)^h \\ \text{----------} \\ (N_m-N_2) & (\varphi_m-\varphi_2) & (\lambda_m-\lambda_2) & (N_m-N_2)^2 & (\varphi_m-\varphi_2)^2 & (\lambda_m-\lambda_2)^2 & \ldots\ldots\ldots & (N_m-N_2)^h & (\varphi_m-\varphi_2)^h & (\lambda_m-\lambda_2)^h \end{pmatrix}$$

$$F_{mm} = \begin{pmatrix} (N_1-N_m) & (\varphi_1-\varphi_m) & (\lambda_1-\lambda_m) & (N_1-N_m)^2 & (\varphi_1-\varphi_m)^2 & (\lambda_1-\lambda_m)^2 & \ldots\ldots\ldots & (N_1-N_m)^h & (\varphi_1-\varphi_m)^h & (\lambda_1-\lambda_m)^h \\ (N_2-N_m) & (\varphi_2-\varphi_m) & (\lambda_2-\lambda_m) & (N_2-N_m)^2 & (\varphi_2-\varphi_m)^2 & (\lambda_2-\lambda_m)^2 & \ldots\ldots\ldots & (N_2-N_m)^h & (\varphi_2-\varphi_m)^h & (\lambda_2-\lambda_m)^h \\ \text{----------} \\ (N_{m-1}-N_m) & (\varphi_{m-1}-\varphi_m) & (\lambda_{m-1}-\lambda_m) & (N_{m-1}-N_m)^2 & (\varphi_{m-1}-\varphi_m)^2 & (\lambda_{m-1}-\lambda_m)^2 & \ldots\ldots\ldots & (N_{m-1}-N_m)^h & (\varphi_{m-1}-\varphi_m)^h & (\lambda_{m-1}-\lambda_m)^h \end{pmatrix}$$

Все внедиагональные блоки матрицы F представляют собой нулевые матрицы размера $(m-1)\times 3h$.

Конкретный вид зависимости (2) в существенной степени зависит от объема результатов проведенных экспедиционных популяционно-генетических исследований (количества обследованных административно-территориальных образований m). Анализ показывает, что в силу объективного наличия временных и материальных ограничений, количество административно-территориальных единиц, охваченных экспедиционными популяционными исследованиями, весьма ограничено. Учитывая это, представляют интерес следующие варианты моделей геногеографического прогноза, непосредственно следующие из общей модели (2):

1) линейная модель геногеографического прогноза с переменными коэффициентами

$$N^*(N,\varphi,\lambda)\approx N^*(N_i,\varphi_i,\lambda_i)+\left(\frac{\partial N^*}{\partial N}\right)_i(N-N_i)+\left(\frac{\partial N^*}{\partial \varphi}\right)_i(\varphi-\varphi_i)+\left(\frac{\partial N^*}{\partial \lambda}\right)_i(\lambda-\lambda_i), \qquad (6)$$

которая применяется, если число административно-территориальных единиц, охваченных экспедиционными популяционно-генетическими исследованиями составляет $4\le m\le 6$.

2) квадратичная модель геногеографического прогноза с переменными коэффициентами

$$N^*(N,\varphi,\lambda)\approx N^*(N_i,\varphi_i,\lambda_i)+\left(\frac{\partial N^*}{\partial N}\right)_i(N-N_i)+\left(\frac{\partial N^*}{\partial \varphi}\right)_i(\varphi-\varphi_i)+\left(\frac{\partial N^*}{\partial \lambda}\right)_i(\lambda-\lambda_i)+$$
$$+\left(\frac{\partial^2 N^*}{\partial N^2}\right)_i(N-N_i)^2+\left(\frac{\partial^2 N^*}{\partial \varphi^2}\right)_i(\varphi-\varphi_i)^2+\left(\frac{\partial^2 N^*}{\partial \lambda^2}\right)_i(\lambda-\lambda_i)^2, \qquad (7)$$

которую целесообразно использовать, если при $7\le m\le 9$.

3) кубическая модель геногеографического прогноза с переменными коэффициентами

$$N^*(N,\varphi,\lambda)\approx N^*(N_i,\varphi_i,\lambda_i)+\left(\frac{\partial N^*}{\partial N}\right)_i(N-N_i)+\left(\frac{\partial N^*}{\partial \varphi}\right)_i(\varphi-\varphi_i)+\left(\frac{\partial N^*}{\partial \lambda}\right)_i(\lambda-\lambda_i)+$$
$$+\left(\frac{\partial^2 N^*}{\partial N^2}\right)_i(N-N_i)^2+\left(\frac{\partial^2 N^*}{\partial \varphi^2}\right)_i(\varphi-\varphi_i)^2+\left(\frac{\partial^2 N^*}{\partial \lambda^2}\right)_i(\lambda-\lambda_i)^2+$$
$$+\left(\frac{\partial^3 N^*}{\partial N^3}\right)_i(N-N_i)^3+\left(\frac{\partial^3 N^*}{\partial \varphi^3}\right)_i(\varphi-\varphi_i)^3+\left(\frac{\partial^3 N^*}{\partial \lambda^3}\right)_i(\lambda-\lambda_i)^3, \qquad (8)$$

в случае, если $m\ge 10$.

Дальнейшее усложнение структуры модели представляется нецелесообразным, так как в практических условиях объем результатов фактически проведенных популяционных исследований, как правило, ограничен указанными значениями.

После того, как выбрана структура модели (в виде (6), (7) или (8)) и получены оптимальные оценки (3) параметров, с ее помощью может быть осуществлен прогноз значений генетических показателей в любом населенном пункте популяции на основе данных о его географической локализации и численности населения. Схему прогноза иллюстрирует рис.1.

Используя, полученные в результате экспедиционных исследований значения $N_i^*=N^*(N_i, \varphi_i, \lambda_i)$, $i=1,...,m$ на основе модели (в виде (6), (7) или (8)) рассчитываются прогнозные значения $N_{iП}^*=N^*(N,\varphi,\lambda)$, $i=1,...,m$ генетического показателя в населенном пункте с координатами φ, λ и численностью населения N.

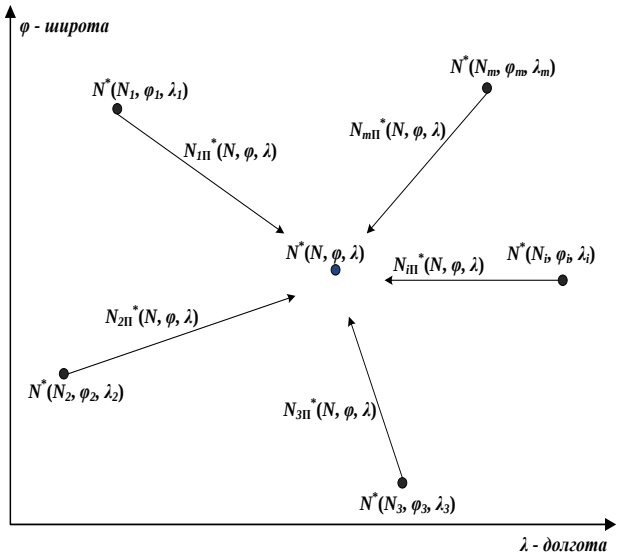

Рис. 1. Иллюстрация метода расчета прогнозируемых значений генетического показателя

На основе совокупности полученных прогнозных значений рассчитывается окончательная оценка значения показателя на основе метода средневзвешенной интерполяции:

$$N^*(N,\varphi,\lambda) = \frac{\sum_{i=1}^{m} w_i(N,\varphi,\lambda) N^*_{i\text{П}}(N,\varphi,\lambda)}{\sum_{i=1}^{m} w_i(N,\varphi,\lambda)} \quad (9)$$

где весовые коэффициенты w_i вычисляются по формуле:

$$w_i = \frac{1}{\left(\sqrt{(N-N_i)^2+(\varphi-\varphi_i)^2+(\lambda-\lambda_i)^2}\right)^{\alpha}}, \quad (10)$$

α является обратной степенью весовой функции и определяет, насколько значения в близких опорных точках сильнее влияют на поведение функции отклика $N^*(N,\varphi,\lambda)$, чем значения в удаленных опорных точках. Выбор параметра α для каждого вида модели (в виде (6), (7) или (8)) был сделан эмпирически на основе сравнения значений генетических показателей, полученных в ходе экспедиционных исследований ряда районов Ростовской и Кировской областей, со значениями тех же генетических показателей, полученных с помощью моделирования. Причем окончательные оценки значений генетических показателей для процедуры сравнения рассчитывались на основе метода средневзвешенной интерполяции, где значение параметра α варьировалось от 1 до 8.

Оптимальными оказались следующие значения параметра α: для линейной модели (6) α=6; для квадратичной модели (7) α=5; для кубической модели (8) α=4.

Как следует из теории оптимального планирования эксперимента точность экспериментальной модели (2) зависит о того, какие именно точки (административно-территориальные единицы) использованы для дальнейшего построения модели. Учитывая это, в диссертационной работе разработан метод формирования оптимального плана экспериментальных исследований, позволяющий определить конкретный набор административно-территориальных единиц, являющихся объектами популяционно-генетических исследований, таким образом, чтобы модель геногеографического прогноза, построенная на основе результатов этих исследований, обладала максимальной точностью. В качестве основы разработанного метода использовался критерий *D*-оптимальности плана эксперимента, поскольку данный критерий обеспечивает сопоставимую точность по сравнению с другими критериями и его вычислительная реализация существенно проще, чем, например, реализация критериев *G*- и *Q*-оптимальности, использование которых приводит к необходимости решения минимаксной задачи.

Задача оптимального планирования эксперимента с целью построения модели геногеографического прогноза рассматривалась в предположении о том, что в границах исследуемой популяции расположено ограниченное число n населенных пунктов с известной географической локализацией φ_j, λ_j и

численностью населения N_j, j=1,..,n, которые рассматриваются в качестве потенциальных объектов генетических исследований. Допустим, что располагаемые материальные ресурсы и временные ограничения позволяют провести экспедиционные генетические исследования в m населенных пунктах. Тогда, применительно, к задаче построения модели геногеографического прогноза матрица плана эксперимента представляет собой матрицу размера mx3 с элементами:

$$X = \begin{bmatrix} \varphi_1 & \lambda_1 & N_1 \\ -- & -- & -- \\ \varphi_m & \lambda_m & N_m \end{bmatrix} \tag{11}$$

Тогда, использование D-оптимального плана приводит к необходимости решения задачи оптимизации следующего вида:

$$|C(X^*)| = min_{X \in W}|C(X)| = min_{X \in W}|(F^T F)^{-1}|, \tag{12}$$

где X* - оптимальный план, W-множество территориально-административных единиц в границах исследуемой популяции, F – матрица (5), конкретный вид которой определяется в зависимости от структуры модели геногеографического прогноза, описываемой выражениями (6)-(8); $C = (F^T F)^{-1}$ – дисперсионная матрица.

В вычислительном плане реализация условия оптимальности (12) приводит к необходимости отыскания минимума неявно заданной скалярной нелинейной функции в пространстве 3m переменных φ_i, λ_i, N_i, i=1,..,m, при наличии ограничений. Однако, учитывая, что область планирования эксперимента W в данном случае представлена дискретным набором значений (φ_j, λ_j, N_j), j=1,...,n, число которых определяется числом населенных пунктов, потенциально пригодных для проведения генетических исследований в границах популяции, решение задачи выбора оптимального плана достигнуто путем перебора всех возможных вариантов сочетаний C_m^n из n по m:

$$C_m^n = \frac{n!}{(n-m)!m!} \tag{13}$$

Для каждого из вариантов сочетаний с учетом структуры модели геногеографического прогноза (в виде (6), (7) или (8)) рассчитывается матрица F и соответствующая ей дисперсионная матрица $C = (F^T F)^{-1}$. План, при котором определитель дисперсионной матрицы C принимает минимальное значение и является D-оптимальным планом, использование которого гарантирует построение модели геногеографического прогноза, обладающей максимальной точностью.

На основе полученных в главе 2 результатов подтверждена целесообразность создания программного комплекса, реализующего описанные выше возможности.

Третья глава диссертационной работы посвящена описанию программного комплекса *GEN*, реализующего разработанную адаптивную модель геногеографического прогноза. Структура программного комплекса *GEN* приведена на рис. 2.

Программный комплекс *GEN* состоит из следующих основных блоков:

1) базы данных результатов популяционно-генетических исследований;

2) блока формирования *D*-оптимального плана экспедиционных исследований;

3) блока адаптации структуры модели и оценки ее параметров;

4) блока прогноза значений генетических показателей в рамках исследуемой популяции, с помощью выбранной модели;

5) блока визуализации результатов, который обеспечивает отображение на географической карте данных, полученных в ходе экспедиционных популяционно-генетических исследований и результатов прогноза, рассчитанных с помощью адаптивной математической модели.

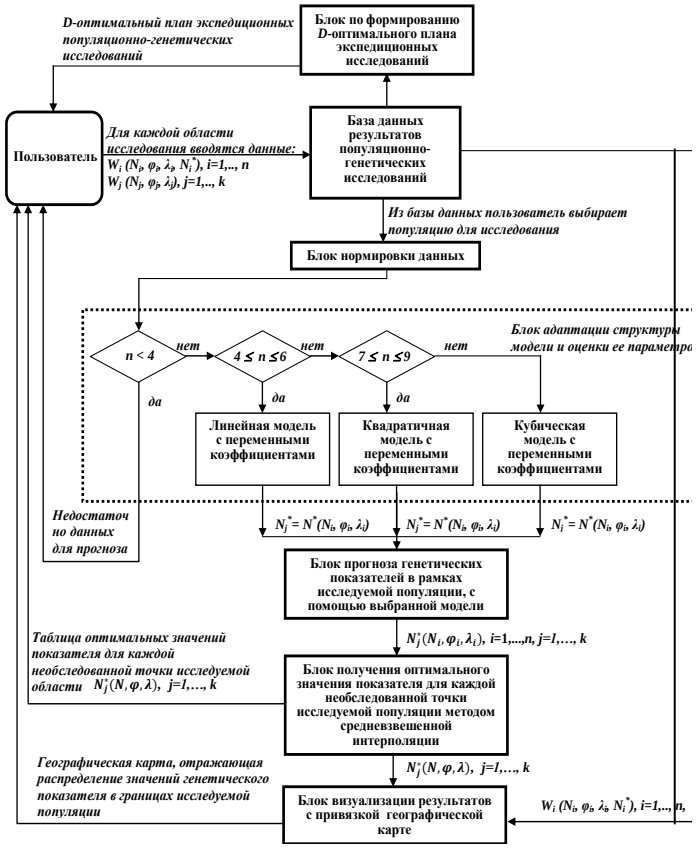

Рис. 2 Структура программного комплекса *GEN*

Выше отмечалось, что в зависимости от объема информации, потенциально доступной для разработки модели, возможны различные варианты ее реализации. Учитывая, что в практических условиях объем результатов фактически проведенных популяционных исследований ограничен, в программном комплексе *GEN* реализованы модели видов (8)-(10) , структура которых автоматически выбирается в зависимости от объема входной информации, за счет чего достигается повышение точности прогноза.

Разработанный программный комплекс *GEN* снабжен простым и наглядным интерфейсом (рис. 3), поддерживающим операции ввода, накопления и хранения данных, который кроме возможности получения прогнозных значений генетических показателей на основе адаптивной модели обладает возможность графической реализации полученных результатов на географической карте исследуемой области (рис. 4).

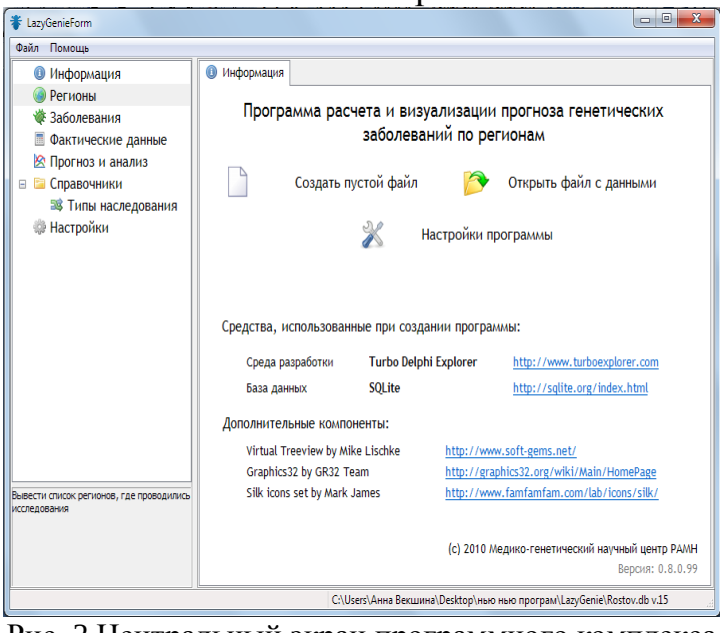

Рис. 3 Центральный экран программного комплекса *GEN*

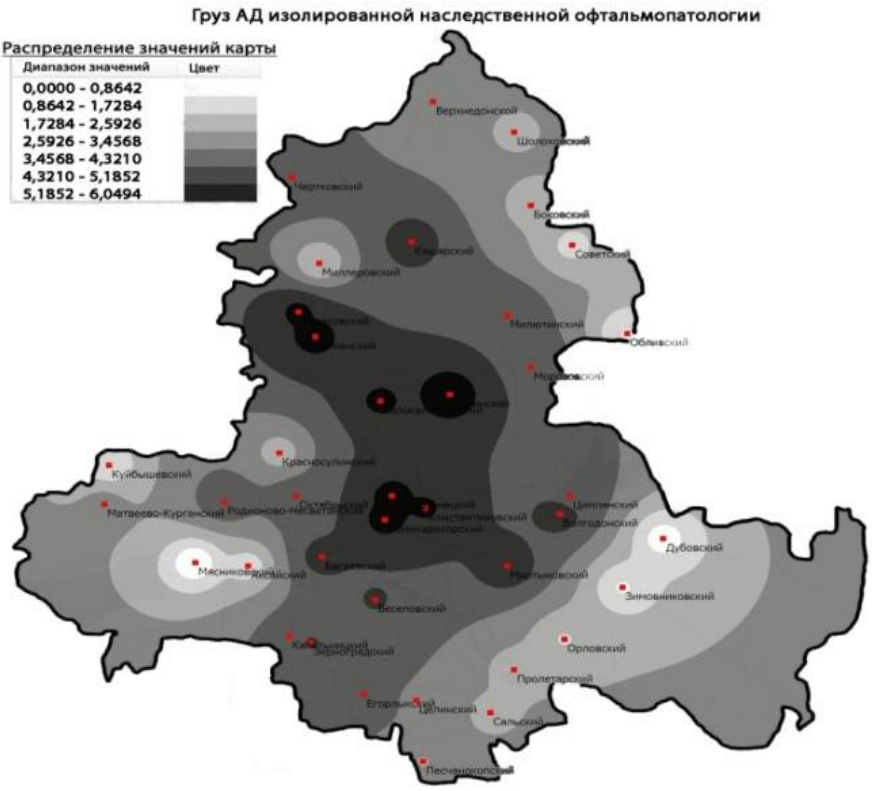

Рис. 4 Груз АД изолированной наследственной офтальмопатологии населения Ростовской области

Четвертая глава диссертационной работы посвящена оценке эффективности разработанной адаптивной модели геногеографического прогноза с использованием экспедиционных популяционно-генетических исследований по аутосомно-рецессивной, аутосомно-доминантной и *X*-сцепленной патологии населения Ростовской и Кировской областей.

Прежде всего, проведено сравнение точности результатов прогноза значений генетических показателей, связанных с наследственными заболеваниями населения различных административно-территориальных образований Ростовской области, полученных с помощью геногеографической модели адаптивной структуры, с оценками, рассчитанными на основе геногеографической модели неизменной линейной структуры. Анализ проводился на основе сравнения значений генетических показателей, полученных в ходе моделирования, с помощью адаптивной и линейной моделей, в ряде выбранных районов Ростовской области, с объективными данными, полученными в ходе экспедиционных популяционно-генетических исследований в этих районах.

Результаты сравнения представлены в виде таблиц и соответствующих им диаграмм (пример одной из диаграмм приведен на рис. 5). Анализ показал, что разработанная адаптивная модель геногеографического прогноза обеспечивает в среднем двукратное повышение точности результатов прогноза по сравнению с геногеографической моделью неизменной линейной структуры.

Так же в главе 4, с помощью разработанного программного комплекса *GEN*, были получены практические результаты прогнозных значений генетических показателей, связанных с наследственными патологиями, в границах Ростовской и Кировской областей в виде таблиц и геногеографических карт.

На основе полученных результатов была проведена проверка адекватности разработанной геногеографической модели путем сравнения двух выборочных совокупностей, объединяющих прогнозируемые с помощью модели и фактически полученные значения показателя отягощенности по некоторому типу наследственного заболевания (пример полученных гистограмм приведен на рис. 6).

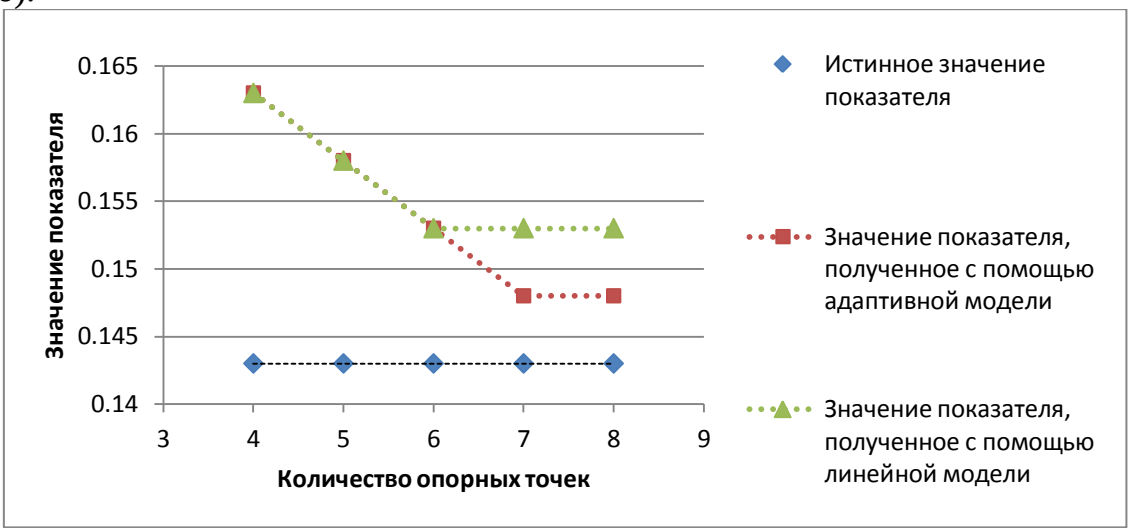

Рис. 5 Распределение значений отягощенности по АР-патологии для Зимовниковского района Ростовской области

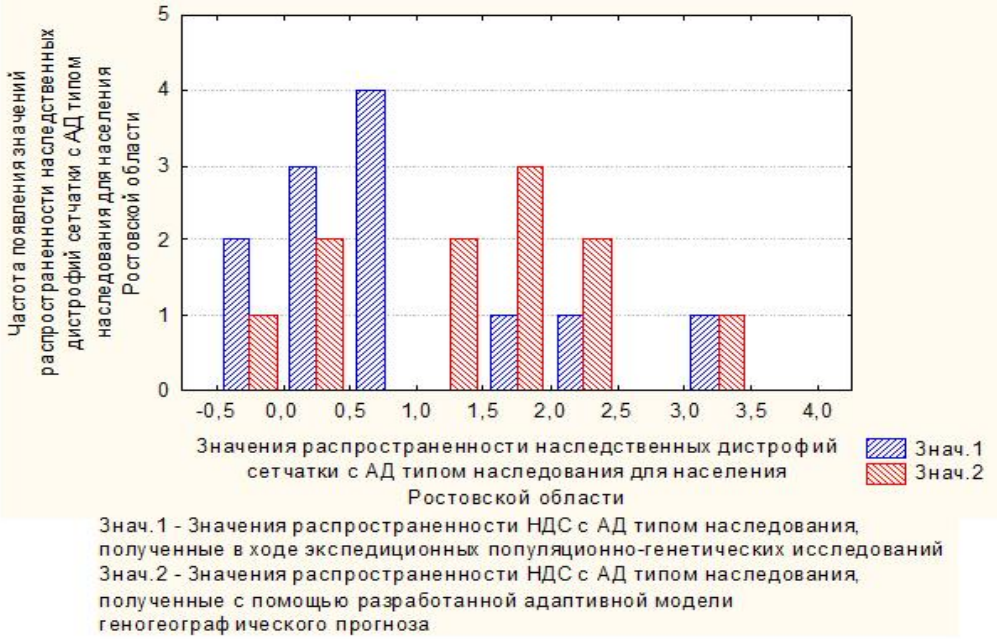

Рис. 6 Гистограммы распространенности наследственных дистрофий сетчатки с АД типом наследования для населения Ростовской области

Примеры результатов сопоставления полученных выборок с использованием двух выборочного критерия Колмогорова-Смирнова приведены в таблицах 1 и 2. Проведенный анализ не выявил статистически значимых различий (на уровне доверительной вероятности 0,95) между прогнозными и фактическими значениями, что подтверждает адекватность модели прогноза.

Таблица 1

Значение статистики Колмогорова-Смирнова для исследований, проведенных в Ростовской области					
	ДС АД	ВК АД	ПРГ АД	ВК АР	ПРГ АР
D^-_{mn}	-0,477273	-0,333333	-0,204545	-0,151515	-0,128788
D^+_{mn}	0,083333	0,166667	0,083333	0,310606	0,234848
p	>0,10	>0,10	>0,10	>0,10	>0,10

Таблица 2

Значение статистики Колмогорова-Смирнова для исследований, проведенных в Кировской области								
	НОП АД	НОП АР	ДС АД	ВК АД	ПРГ АД	ПЗ АД	ДС АР	ВК АР
D^-_{mn}	-0,350000	-0,308333	-0,40000	0,466667	-0,17500	-0,2333	-0,4000	-0,3333
D^+_{mn}	0,275000	0,166667	0,40000	0,066667	0,116667	0,266607	0,2750	0,2250
p	>0,10	>0,10	>0,10	>0,10	>0,10	>0,10	>0,10	>0,10

С целью оценки точности разработанной адаптивной модели геногеографического прогноза было проведено сравнение результатов прогноза значений генетических показателей, связанных с наследственными заболеваниями населения различных районов Ростовской и Кировской областей, полученных с помощью разработанной модели с объективными данными, полученными в ходе экспедиционных популяционно-генетических исследований в этих районах. Были получены диаграммы (пример диаграммы приведен на рис. 7) и табличные значения генетических показателей, анализ которых показал, что в 58% случаев ошибка не превышает 5%-тного уровня от объективного значения генетического показателя и в 87,5% случаев ошибка не превышает 10%-тного уровня.

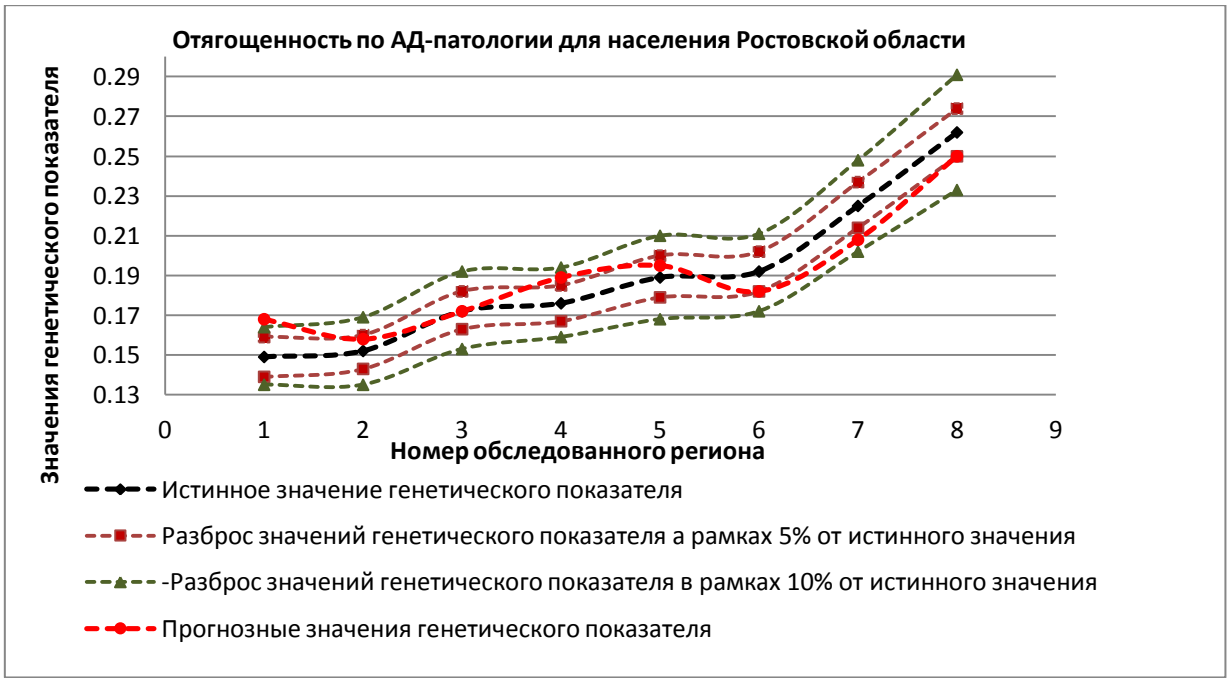

Рис. 7 График распределения значений отягощенности АД-патологией населения Ростовской области в рамках заданной точности

С помощью разработанного метода планирования программы проведения экспедиционных исследований были получены оптимальные планы

популяционно-генетических исследований для Ростовской области, Чувашской и Удмуртской Республик. На рис. 8 приведена зависимость, отражающая изменение значений определителя дисперсионной матрицы *C* при разработке вариантов планирования для Удмуртской Республики.

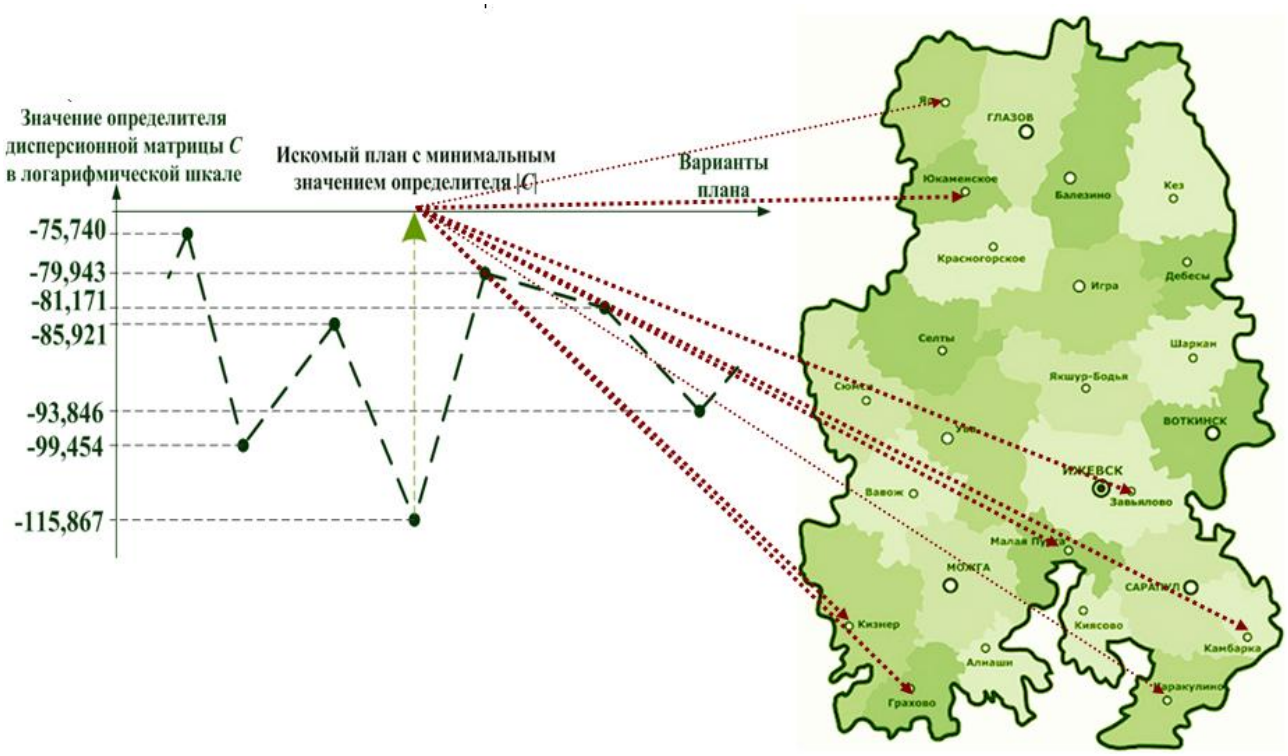

Рис. 8 Иллюстрация процесса определения оптимального плана экспедиционных исследований Удмуртской Республики (*n*=25, *m*=8)

В таблице 3 представлены окончательные результаты построенного плана экспедиционных популяционно-генетических исследований для сельских поселений Удмуртской Республики.

Таблица 3

Количество опорных точек (выбирает исследователь)	План экспедиционных популяционно-генетических исследований для сельских поселений Удмуртской республики (указаны районы)	Минимальные значения критерия *D*-оптимальности (в логарифмической шкале)
4	Граховский, Завьяловский, Каракулинский, Ярский	-22,125
5	Граховский, Завьяловский, Камбарский, Каракулинский, Ярский	-29,119
6	Граховский, Завьяловский, Камбарский, Каракулинский, Кизнерский, Ярский	-36,614
7	Граховский, Завьяловский, Камбарский, Каракулинский, Кизнерский, Юкаменский, Ярский	-89,041
8	Граховский, Завьяловский, Камбарский, Каракулинский, Кизнерский, Малопургинский, Юкаменский, Ярский	-115,867
9	Балезинский, Граховский, Завьяловский, Камбарский, Каракулинский, Кизнерский, Малопургинский, Юкаменский, Ярский	-134,082
10	Балезинский, Граховский, Завьяловский, Камбарский, Каракулинский, Кизнерский, Киясовский,Малопургинский, Юкаменский, Ярский	-246,887

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. На основе анализа и оценки современных методов получения значений генетических показателей в границах исследуемых популяций, показана актуальность и практическая значимость разработки адаптивной математической модели геногеографического прогноза.

2. Разработана адаптивная модель прогноза значений генетических показателей, отличающаяся от существующих аналогов возможностью автоматической реконфигурации ее структуры в зависимости от объема доступных для анализа результатов экспедиционных популяционно-генетических исследований. Разработанная адаптивная модель геногеографического прогноза, обеспечивает повышение точности прогноза значений генетических показателей, связанных с распространением наследственных заболеваний в пределах изучаемой популяции, вследствие включения в ее структуру не только географических координат исследуемых населенных пунктов, но и численности проживающего в них населения.

3. Разработан комплекс алгоритмов, обеспечивающих автоматическую адаптацию и расчет параметров моделей геногеографического прогноза на основе результатов популяционно-генетических исследований;

4. Предложен метод и алгоритм построения *D*-оптимального плана для выбора административно-территориальных единиц, являющихся объектами популяционно-генетических исследований, который обеспечивает получение модели геногеографического прогноза обладающей максимальной точностью.

5. Создан программный комплекс *GEN*, который реализует возможность планирования экспедиционных популяционно-генетических исследований, обеспечивает получение оценок генетических показателей в границах исследуемых популяций на основе разработанных моделей геногеографического прогноза и графическое представление полученных результатов на географической карте.

6. Получены практические оценки показателей отягощенности различными типами наследственных патологий Ростовской и Кировской областей, а так же составлены оптимальные планы экспедиционных популяционно-генетических исследований для Чувашской и Удмуртской Республик.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях из рекомендованного ВАК Минобрнауки России перечня:

1. Векшина А.Б., Евдокименков В.Н., Зинченко Р.А. Компьютерная модель геногеографического прогноза // Вестник компьютерных и информационных технологий. 2011. №12. Стр. 39-47.

2. Ельчинова Г.И., Игумнов П.С., Векшина А.Б., Зинченко Р.А. Инбридинг и эндогамия в Татарстане // Генетика. 2012. Т. 48. № 3. Стр.408-411.

3. Векшина А.Б., Евдокименков В.Н., Зинченко Р.А. Применение методов планирования эксперимента для повышения точности модели геногеографического прогноза // Научное обозрение. 2012. № 2. Стр. 104-108.

Публикации в зарубежных изданиях:

4. A.B. Vekshina, R.A. Zinchenko, V.N. Evdokimenkov, T. Mamedov. The mathematical model of genetic targets in a limited amount of their measurements // European J. of Hum.Gen. - 2009. - V.17, supp.2. - P.223

Другие публикации:

5. Векшина А.Б., Евдокименков В.Н., Зинченко Р.А., Ельчинова Г.И., Игумнов П.С. Математическая модель прогноза значений генетических показателей // Материалы I Междунар. научно-практ. конф.: «Достижения, инновационные направления, перспективы развития и проблемы современной медицинской науки, генетики и биотехнологий», М.: Издательство «Буки Веди» 2011 г. - С. 64-65.

6. Векшина А.Б., Евдокименков В.Н., Зинченко Р.А., Ельчинова Г.И., Игумнов П.С. Математическая модель геногеографического прогноза, построенная на основе методов оптимального оценивания // Труды XI Междунар. научно-практ. конф. «Интеллект и наука», Красноярск: Центр информации, 2011. – С. 209.

7. Векшина А.Б., Евдокименков В.Н., Зинченко Р.А., Ельчинова Г.И., Игумнов П.С. Математическая модель геногеографического прогноза // Эл. издание «Биомедицинская инженерия и биотехнология: сборник материалов IV Всерос. научно-практ. конф. с междунар. участием», ГОУ ВПО «Курский государственный медицинский университет», эл. издание № 23560, стр. 28-29.

8. Векшина А.Б., Суйкова Т.А., Пичулин В.С. Математическое моделирование вентиляционного костюма // Тезисы докладов X Междунар. конф. «Системный анализ, управление и навигация» - М.: Издательство МАИ, 2005 – С. 178-179.

9. Векшина А.Б., Суйкова Т.А., Юров И.Б., Белозерова И.Н., Строгонова Л.Б. Измерение концентрации лактата в капиллярной крови при дозированной физической нагрузке // Тезисы докладов IV Междунар. конф. «Авиация и космонавтика – 2005» - М.: Издательство МАИ, 2005 – С. 105.

Подписано в печать: 20.04.12
Тираж: 100 экз. Заказ № 375
Отпечатано в типографии «Реглет»
119526, г. Москва, ул. Фридриха Энгельса, д.3/5, стр. 2
(495) 661-60-89; www.reglet.ru