

С. М. Кошель

Географический факультет МГУ им. М.В.Ломоносова,
кафедра картографии и геоинформатики
skoshel@geogr.msu.ru

ГЕОИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ГЕНОГЕОГРАФИИ

Геногеография изучает пространственное распределение генофонда популяций – обособленных групп населения, исторически сложившихся на определенной территории и воспроизводящих себя в границах этого «исторического» ареала из поколения в поколение. Географическое картографирование генофонда – единственный способ своими глазами увидеть его зримый образ на картах, проанализировать процессы, протекающие в генофонде.

На рубеже 1980-х годов благодаря органическому соединению с картографией и созданию компьютерных банков данных о частотах генов в населении, начался принципиально новый этап в развитии геногеографии – компьютерное генетическое картографирование. Это означает, что появилась практическая возможность от географии генов человека перейти к географии генофондов населения мира, регионов, отдельных стран и этнических групп, исследовать генофонд не только общими генетико-статистическими методами, но и собственными уникальными методами геногеографии – картографическими.

История развития и применения программного обеспечения для целей геногеографии хорошо описана в монографии (Балановская и др., 2007). За рубежом наиболее значимые работы в этой области выполнялись коллективом под руководством одного из ведущих мировых специалистов по популяционной генетике Кавалли-Сфорца (Piazza et al., 1981). Похожие, но менее мощные программы разрабатывались и другими коллективами популяционных генетиков, например (Sokal et al., 1999). Одной из первых (и остающейся единственной по сей день) серьезных разработок в нашей стране был пакет GG MAG, созданный сотрудниками лаборатории автоматизации кафедры картографии и геоинформатики совместно со специалистами института общей генетики им. Н. И. Вавилова РАН Рычковым Ю. Г. и Балановской Е. В. в начале 1990-х годов (Балановская и др., 1994а, 1994б). В своей основе это был пакет MAG (Сербенюк и др., 1991; Кошель, 2000), дополненный специфическими методами моделирования и статистического анализа. С его помощью было выполнено множество исследований, созданы тысячи карт и десятки атласов, обширный библиографический список этих работ представлен в (Балановская и др., 2007). Однако в

связи с тем, что пакет GG MAG разрабатывался в операционной среде MS DOS, его использование в настоящее время сталкивается с большими трудностями. Кроме того, за почти 20 лет эксплуатации пакет морально устарел – произошли кардинальные изменения в возможностях организации интерфейса пользователя, машинной графике и т. д. Анализ множества российских и зарубежных работ в области популяционной генетики, затрагивающих тематику геногеографии (именно в части создания карт и пространственного моделирования и анализа), показывает, что едва ли не единственным используемым современным программным средством является пакет Surfer, а коммерческие ГИС-пакеты специалистами-генетиками практически не используются. По всей видимости, это связано с высокой стоимостью коммерческих ГИС и большими сложностями в их освоении для неспециалистов. Кроме того, стандартных встроенных средств анализа в коммерческих ГИС явно не хватает и требуется создавать дополнительные программные модули, пользуясь средой программирования ГИС-пакетов. В связи с этим назрела необходимость разработки нового программного обеспечения, использующего современный арсенал геоинформационных технологий и, с одной стороны, максимально приспособленного для нужд геногеографии (в отличие от пакета Surfer), а с другой стороны, обеспечивающего простоту создания и высокое качество карт при минимальных требованиях к знаниям пользователя в области картографии и геоинформатики (в отличие от коммерческих ГИС-пакетов).

В настоящее время в лаборатории автоматизации кафедры картографии и геоинформатики совместно со специалистами из лаборатории популяционной генетики МГНЦ РАН (Балановский О. П., Балановская Е. В.) разрабатывается новая версия пакета GG MAG, получившая название GeneGeo. Её основными особенностями являются: возможность создания цифровых моделей генетических признаков с использованием сферических расстояний между точками; наличие подготовленных заранее картографических основ на различные регионы и мир в целом и автоматическое создание легенды карт и дополнительной статистической информации о признаке, что позволяет комфортно работать с программой не специалистам в области картографии; включение специализированных картографо-статистических методов для анализа данных (тренды, корреляции и т.д.) и построения «синтетических» карт (карт изменчивости совокупности генетических признаков – главных компонент, общего генетического разнообразия, межпопуляционных генетических различий, гетерозиготности, генетических расстояний и т.д.). Остановимся более под-

робно на описании трех основных блоков программы.

Блок *моделирования* предназначен для создания сеточных цифровых моделей признаков по их числовым значениям в опорных точках. Как правило, в роли признаков выступают частоты так называемых генетических маркёров. К ним относятся (Балановская и др., 2007) классические генетические маркёры (физиологические, иммунологические, биохимические), ДНК маркёры (диаллельные и мультиаллельные, аутосомные и «однородительские»), квазигенетические маркёры (фамилии, демографические параметры и др.). Координаты опорных точек задаются в широте и долготе, программа их автоматически пересчитывает в проекцию текущей выбранной картографической основы и цифровая модель создается на сетке в этой же проекции. Опыт эксплуатации пакета GG MAG показал, что из всего многообразия методов моделирования для перечисленных выше статистических геополей наиболее подходящим является метод Шепарда – обобщение метода обратных взвешенных расстояний (Сербенюк и др., 1991), поэтому именно он был реализован в первую очередь. За счет множества параметров (степень весовой функции, вид аппроксимирующей функции в опорных точках, параметр сглаживания, радиус влияния и др.) достигается большая вариативность метода, возможность подобрать свой наиболее подходящий вариант для каждого из признаков. По возможностям настройки метод существенно превосходит аналогичные реализации в других паке-

тах. Еще одной отличительной (и уникальной, разработанной специально для GeneGeo) особенностью является вычисление расстояний между точками, используемых в интерполяционных формулах, не в декартовых координатах проекции или пространства «долгота-широта», а на сфере. Это позволяет корректно строить модели континентального и глобального (весь мир) масштаба без влияния искажения длин в проекции и проблемы скачкообразного изменения значений долготы на 2π при переходе через определенный меридиан. Созданные цифровые модели сохраняются в виде файла в специально разработанном формате, в котором, помимо матрицы значений признака в узлах регулярной сетки, содержится информация об исходных данных, параметрах метода моделирования, способе визуализации модели и основная статистика. Файл цифровой модели в дальнейшем может быть загружен в программу как сеточный слой. Качественное и объективное исследование популяции геногеографическими методами предполагает изучение сотен признаков, поэтому в блоке моделирования предусмотрен пакетный режим, позволяющий одновременно создавать модели сразу для всех признаков, содержащихся в таблице с исходными данными. В дальнейших планах по развитию блока моделирования значится реализация метода кригинга и аппроксимации с помощью сферических ортогональных многочленов.

Блок *пространственного и статистического анализа* наиболее обширен и составляет «сердце» программы. Исходными данными для анализа



Рис. 1. Макет карты на мир в тройной проекции Винкеля с вариантом расположения элементов зарамочного оформления

являются цифровые модели, результатом – тоже, как правило, цифровая модель статистического («синтетического») показателя. Уже реализован модуль для вычисления разнообразных статистик методом «скользящего» окна (размер и форма окна выбирается пользователем), как стандартных (среднее, размах, дисперсия и др.), так и специфических (характеристики общего, внутри- и межпопуляционного генного разнообразия, частных коэффициентов корреляции с широтой или долготой и др.). Здесь же имеется возможность вычисления цифровых моделей линейных (Пирсона) или ранговых (Кендэлла) коэффициентов корреляции между двумя признаками. Тренд-анализ (построение трендов и остаточных моделей) пока реализован только в виде алгебраических многочленов, в дальнейшем предполагается использовать ортогональные сферические многочлены в качестве базисных функций тренда.

Поскольку заранее трудно предусмотреть все многообразие возможных расчетов с цифровыми моделями, в блоке анализа реализован сеточный калькулятор, в котором пользователь может задать математическую формулу для вычислений и указать цифровые модели, соответствующие переменным в этой формуле (можно использовать до трех переменных). Однако далеко не все расчеты могут быть выполнены с помощью такого «скалярного» калькулятора, разрабатывается также «векторный» калькулятор, в котором в качестве переменной будет выступать не одна модель, а вектор моделей и будут присутствовать специфические для такого рода данных функции (например, скалярное произведение, сумма всех компонент вектора и т. д.).

Наиболее часто используемые, имеющие свои названия, методы анализа вынесены в отдельные модули, хотя некоторые из них могут быть реализованы и в рамках скалярного или векторного калькулятора. В основном это создание «синтетических» или обобщенных моделей (а затем и карт) изменчивости совокупности признаков, служащих индикатором состояния генофонда. Сюда входит уже реализованный расчет генетических расстояний до реперных популяций по Нею (Nei, 1975), уровня гетерозиготности, общего генетического разнообразия, межпопуляционных генетических различий и др., а также ряд методов многомерного статистического анализа (метод главных компонент, многомерное шкалирование, кластерный анализ и т. д.). Созданные обобщенные статистические модели также сохраняются в виде файла цифровой модели и затем могут быть загружены в картографический блок как сеточный слой.

Картографический блок организован, исходя из идей, изложенных выше, а именно, он должен обеспечивать максимальное качество создаваемых карт, использовать современные геоинформаци-

онные технологии и при этом не требовать от пользователя каких-либо специфических познаний и высокой образованности в области картографии и геоинформатики. Поэтому картографические основы карт на различные территории, от регионального до глобального уровня (включая макеты оформления), созданы заранее, пользователю нужно только выбрать основу на ту территорию, с которой он будет в настоящий момент работать. Практически для всех карт используется равновеликая азимутальная проекция Ламберта. Такой выбор диктовался тем, что в качестве возможного элемента оформления карты фигурирует гистограмма распределения показателя, вычисляемая по его цифровой модели «на лету». В случае равновеликой проекции расчет гистограммы значительно упрощается, азимутальная же проекция позволяет получить изображение земной поверхности с меньшими искажениями форм. Для карт материков использованы параметры проекций, предлагаемые в учебнике (Серапинас, 2005). Единственным исключением является основа на весь мир, она выполнена в тройной проекции Винкеля (два варианта с различными базовыми меридианами), а расчет гистограммы в этом случае выполняется по отдельному алгоритму (вариант макета карты на мир с базовым меридианом 150° в.д. показан на рис. 1). Масштаб карт выбирается таким образом, чтобы карта размещалась при печати на листе А4, но в основу закладывается небольшая избыточная детальность, чтобы при необходимости ее можно было увеличить до размеров формата А3. Поскольку пользователю предоставлена возможность получать изображение произвольного размера (в том числе при экспорте карты в растровый формат), на карте показывается только масштабная линейка без указания численного или именованного масштаба.

Основа карты включает следующие базовые слои: береговая линия, границы государств, столицы и некоторые города, главные реки и озера, второстепенные реки и озера, рельеф суши (сеточная цифровая модель). Градусная сетка с подписями рисуется автоматически с заданным пользователем шагом по широте и долготе (возможно различным). При этом пользователь может по своему усмотрению включать и выключать отображение того или иного слоя на карте. В макет оформления карты также включены окна для размещения названия карты, легенды, масштабной линейки и статистической информации, в том числе гистограммы распределения показателя, как по значениям цифровой модели, так и по данным в опорных точках.

Для отображения цифровых моделей предусмотрены способы изолиний и послойной окраски с возможностью совмещения послойной окраски с аналитической отмывкой. Специально для

программы GeneGeo был разработан новый алгоритм совмещения аналитической отмычки с цветовым фоном карты, в котором яркость изображения не только уменьшается в зависимости от угла между нормалью к поверхности и направлением на источник освещения (как это делалось до сих пор), но и увеличивается при переходе через некоторое заданное пороговое значение угла. Пороговое значение подбирается так, чтобы для плоских участков рельефа цвет фона не изменялся (это зависит от угловой высоты источника освещения над горизонтом), при этом яркость увеличивается при значениях угла меньше порогового значения. В результате цвет на карте воспринимается визуально так же, как он выглядит в легенде, в отличие от стандартного подхода, при котором цвет на карте с отмычкой выглядит более темным и «тусклым». Существенным облегчением при создании цветовой шкалы для способа по-

слоиной окраски является возможность интерполяции цвета между заданными фиксированными значениями, то есть, цвет задается пользователем только для некоторых интервалов шкалы, цвета же остальных интервалов интерполируются программой, обеспечивая плавный переход от одного фиксированного цвета к другому. Реализована также возможность сохранения числовой и цветовой шкалы в файл с последующей загрузкой, что очень удобно при создании серии карт в единой шкале.

Таким образом, при создании карты от пользователя требуется только выбрать способ и параметры отображения цифровой модели картографируемого признака, указать нужные элементы оформления из предлагаемого набора и ввести текст заголовка и легенды, все остальное программа делает автоматически. Отметим, что речь идет об электронных картах, где изображение

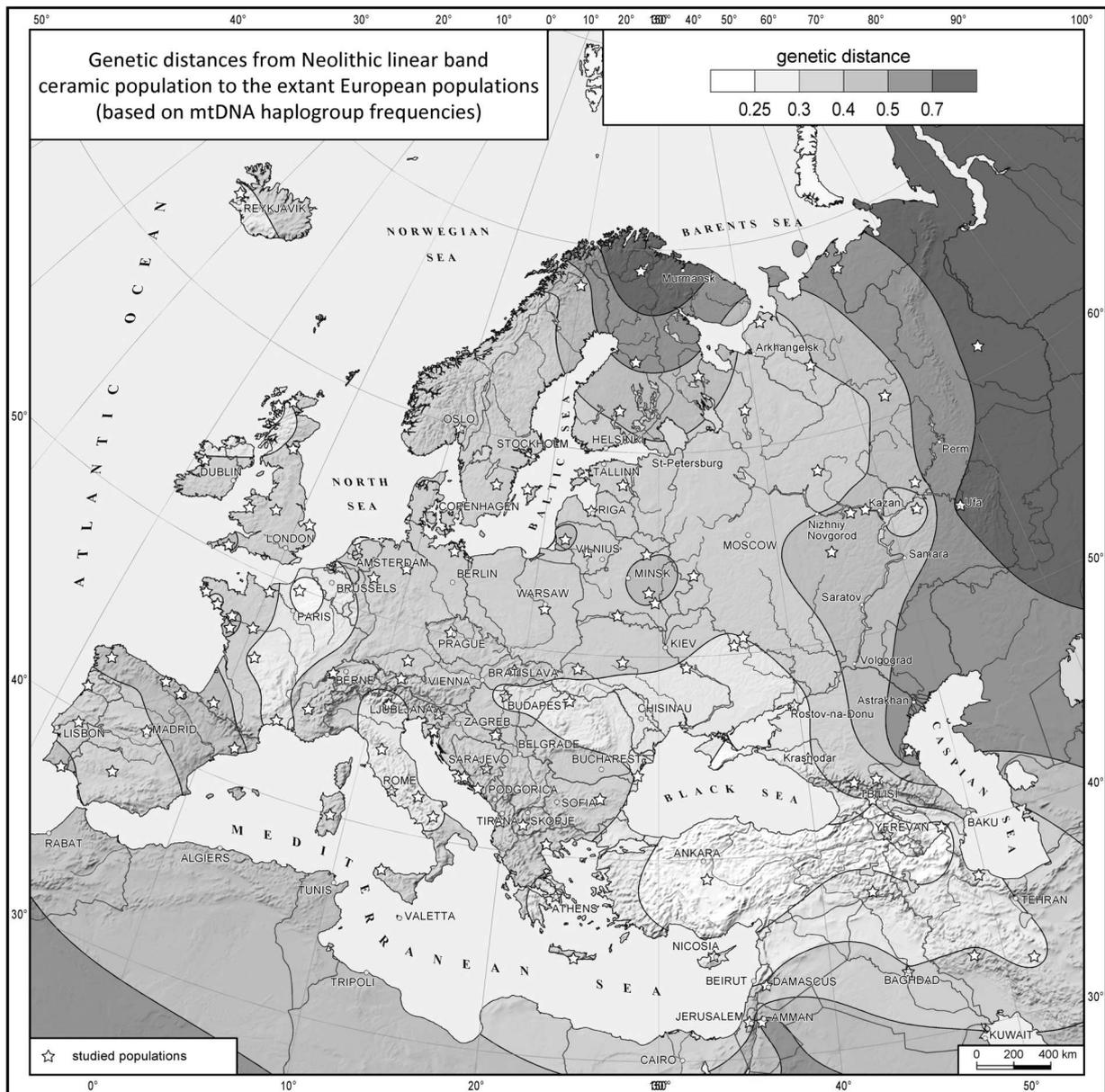


Рис. 2. Карта генетических расстояний от древних неолитических образцов (ЛБК – период культуры линейно-ленточной керамики) до современной популяции Европы и Ближнего Востока

непосредственно связано с цифровой моделью показателя и изменяется в режиме реального времени при изменении параметров отображения. Кроме того, полученную карту можно вывести на печать или сохранить в растровом формате.

Разработка пакета GeneGeo была начата в 2008 году, а уже в 2009 году с его помощью были созданы первые карты. К настоящему времени опубликован целый ряд научных работ (в том числе и в зарубежных изданиях), в которых пакет GeneGeo использован как для картографирования, так и для анализа данных. В частности, ряд интересных карт генетических расстояний от древних неолитических образцов населения периода культуры линейно-ленточной керамики (ЛБК) до всей современной популяции Европы и Ближнего Востока представлен в статье (Haak et al., 2010). Адаптированный черно-белый вариант одной из таких карт показан на рис. 2. С некоторыми сериями карт можно ознакомиться (а заодно и сравнить с картами, созданными в старой версии GGMAG) на сайте лаборатории популяционной генетики МГНЦ РАН (www.genofond.ru, раздел «Атласы»).

Литература

1. Nei M. Molecular population genetics and evolution. – Amsterdam: North-Holland Publ. Co., 1975. – 290 p.
2. Piazza A., Menozzi P., Cavalli-Sforza L. L. The making and testing of geographic gene frequency maps // *Biometrics*, 1981, v.37, p.635-659.
3. Сербенюк С. Н., Кошель С. М., Мусин О. Р. Программы МАГ для создания цифровых моделей геополей // *Геодезия и картография*, 1991, № 4, с.44-46.
4. Балановская Е. В., Нурбаев С. Д., Рычков Ю. Г. Компьютерная технология геногеографического изучения генофонда. I. Статистическая информация карт // *Генетика*, 1994а, Т.30, №7, с.951-965.
5. Балановская Е. В., Нурбаев С. Д., Рычков Ю. Г. Компьютерная технология геногеографического изучения генофонда. II. Статистическая трансформация карт // *Генетика*, 1994б, Т.30, №11, с.1538-1555.
6. Sokal R. R., Oden N. L., Thomson B. A. A problem with synthetic map // *Human Biology*, 1999, v.71, N1, p.1-13
7. Кошель С. М. Цифровое моделирование и анализ геополей с помощью пакета "МАГ" // *Взаимодействие картографии и геоинформатики* (ред. А. М. Берлянт, О. Р. Мусин). – М.: Научный мир, 2000. – с.41-49.
8. Серапинас Б. Б. Математическая картография: Учебник для вузов. – М.: Издательский центр «Академия», 2005. – 336 с.
9. Балановская Е. В., Балановский О. П. Русский генофонд на Русской равнине. – М.: ООО «Луч», 2007. – 416 с.
10. Haak W., Balanovsky O., Sanchez J. J., Koshel S., Zaporozhchenko V., et al. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities // *PLoS Biology*, 2010, 8(11), p.1-16: e1000536. doi:10.1371/journal.pbio.1000536.