Картографирование данных рейтингов «Политического Атласа Современности»

Резюме
В рамках проекта «Политический Атлас Современности» (http://worldpolities.org/), разработанным совместно МГИМО, Институтом Общественного Проектирования и журналом «Эксперт», в журнале «Полис» были опубликованы данные рейтингов 192 стран мира по пяти специально разработанным интегральным индексам (уровня жизни, международного влияния, угроз, государственности и демократии). Дополняя анализ данных, проделанный в [1], мы рассматриваем различные методологические аспекты статистического анализа «Атласа» и предлагаем разработанный лабораторией нейроинформатики Института Вычислительного Моделирования СО РАН (г.Красноярск) подход - картографирование данных, который ранее успешно применялся для анализа экономических, социологических и биологических данных. Особенностью предлагаемого метода является его способность выявлять нелинейные закономерности в наборе данных и наглядно представлять многомерную информацию в виде карты данных и сопутствующего атласа информационных раскрасок. Полученные результаты проецируются на географическую карту мира. Результаты анализа, представленные в виде интерактивной веб-страницы, доступны на http://atlas.savvy.ru .

Введение

Проект «Политический Атлас Современности», разрабатываемый совместно МГИМО, Институтом общественного проектирования и журналом «Эксперт», представляет из себя уникальную попытку провести многомерный анализ закономерностей и тендеций существующих во всем разнообразий политических устройств в современном мире.

Для анализа были разработаны около 60 индикаторов, по которым сравниваются 192 страны мира. Индикаторы были агрегированы в 5 индексов, позволяющих дать обобщенную оценку того или иного государства. Это индексы уровня жизни, международного влияния, внешних и внутренних угроз, государственности и демократии. Методика построения индексов подробна изложена в [1] и основана на построении обучающих выборок «типичных» для индекса государств (государства с развитой и неразвитой демократией, влиятельные и невлиятельные державы, и т.д.) и обобщении этого экпертного знания с помощью методологии линейного дискриминантного анализа.

В работе [1] был проведен подробный анализ данных с применением метода главных компонент и кластерного анализа. Сделанные выводы имеют большое значение для понимания ... В частности, проведенный анализ показывает, что ...

В настоящей работе, дополняя результаты [1], мы проводим независимый анализ данных рейтингов Политического Атласа Современности и применяем разработанную технологию *картографирования данных*, основанную на построении малоразмерного нелинейного главного многообразия, вложенного в многомерное пространство данных. Эта технология была разработана несколько лет назад лабораторией нейроинформатики Института Вычислительного Моделирования СО РАН (г.Красноярск) к.т.н. А.Ю.Зиновьевым и к.т.н. А.А.Питенко под руководством д.ф.-м.н., проф. А.Н.Горбаня. Лежащий в основе *метод упругих карт* обобщает линейный анализ главных компонент, позволяет выявить и использовать нелинейности в геометрической структуре облака данных и представить результаты анализа в виде двумерной карты данных и набора (атласа) информационных раскрасок, наложенных на карту. Математические принципы построения карты данных изложены в работах [2-5]. Технология воплощена в свободно распространяемом программном продукте VidaExpert, доступном с вебсайта авторов программы. В настоящее время метод развивается и применяется на практике в сотрудничестве с Институтом Кюри (Париж, Франция) и Центром Математического Моделироания Лейстерского Университета (Англия). Метод картографирования данных ранее успешно применялся в анализе экономических, социологических и биологических данных [2-6].

Данная работа построена следующим образом. В первом разделе мы представляем карту данных и информационный атлас раскрасок, представляющий структуру распределения государств в пятимерном пространстве индексов. Второй раздел посвящен проекции результатов анализа на мировую географическую карту. В третьем разделе мы рассматриваем некоторые методологические аспекты

проведенной работы в сравнении с уже проведенным статистическим анализом, а также предлагаем некоторые направления дальнейших исследований.

1. Карта данных индексов

Идея картографирования данных заключается в представлении таблицы данных на плоскости или в трехмерном пространстве таким образом, чтобы объекты (строки таблицы), обладающие сходными характеристиками (значениями столбцов в таблице) оказались по возможности по соседству с друг другом. С возможностями и различными вариантами такого подхода можно познакомиться, прочитав, например, монографию [2] .

Карта данных в нашей работе представляет из себя нелинейный двумерный экран, называемый *упругой картой*, вложенный в пятимерное пространство индексов. Этот экран позволяет осуществить проекцию таблицы данных индексов, представленную как «облако точек» в пространстве индексов, из пятимерия на двумерную плоскость. На языке математики, упругая карта является аппроксимацией *главного многообразия* для набора точек, то есть, с одной стороны, минимизирует средний квадрат расстояния от точек до карты, с другой – обладает свойствами *гладкости*, то есть двумерная поверхность, будучи вложенной в пятимерие, не слишком сильно изогнута и перекручена. Метод упругих карт [3] естественным образом обобщает одновременно *метод главных компонент* [7], широко применяемый при статистическом анализе данных, и *метод динамических ядер* (более известный как K-means в анлоязычной литературе), также широко применяемый для кластеризации данных [7].

Некоторое представление о форме нелинейного экрана, вложенного в пятимерие, можно получить из рис.1 (внизу справа), где показана проекция построенного многообразия в трехмерном пространстве, натянутом на первые три главные компоненты распределения точек данных.

На рис.1 показана карта данных индексов Политического Атласа Современности и сопутствующий атлас информационных раскрасок. Цветом и формой отмечены различные группы стран. Большинство групп стран образовано по географическому положению, но нами было решено выделить в отдельную группу страны постсоветского пространства, а также отдельным символом выделена Россия. Размер точки на карте пропорционален логарифму населения страны.

По построению, близкие точки соответствуют странам с похожими значениями индексов. На раскраске на рис.1 вверху справа цветом изображена плотность точек на карте, что позволяет выявить кластеры или «сгущения» в распределении данных. Эта раскраска показывает, что в распределении точек нет выраженной кластерной структуры и следует с осторожностью говорить о «кластерах стран». Тем не менее можно выделить порядка десятка «локальных сгущений», объединяющих страны с похожими профилями индексов. Так, можно выделить следующие «сгущения» :

K05) большое *сгущение «типичных» стран* (другими словами, стран из «серой зоны»), объединяющий страны со значениями индексов близких к средним; здесь мы обнаруживаем некоторые страны Центральной и Южной Америки (Панама, Гватемала, Коста Рика, Парагвай, Суринам, Эквадор, Гондурас), крупные страны Карибов (Ямайка, Доминиканская Республика, Тринидад и Тобаго, менее типичны Багамы), Маврикий, Ботсвана, Хорватия; сюда же попали Латвия, Литва и Украина; к этому же кластеру примыкает *кластер стран Восточной Европы* (L07) и Уругвай.

O09) *европейский кластер*, где основная вариативность идет по индексу международного влияния; в порядке увеличения этого индекса - Ирландия, Финляндия, Дания, Норвегия, Австрия, Швеция, Швейцария (сдвинута по градиенту уровня жизни), Бельгия, Нидерланды; в этом же кластере естественно оказываются Новая Зеландия, Австралия и Канада; Португалия несколько выпадает из этого сгущения и оказывается ближе к сгущению стран Восточной Европы. Испания и Греция, напротив, оказываются рядом со сгущением развитых стран Южной и Центральной Америки.

P13) *европейская большая четверка*, в порядке увеличения индекса влияния – Италия, Великобритания, Франция, Германия.

P02) сгущение *маленьких стран Западной и Южной Европы* (Монако, Лихтенштейн, Андорра, Сан Марино). Люксембург и Исландия не входят в это сгущение и стоят особняком.

J09) *сгущение «типичных» тоталитарных государств* (низкий индекс демократичности и средние значения остальных индексов): это многие страны Среднего Востока и Азии (Оман, ОАЭ, Кувейт, Бахрайн, Бруней, Катар, Йордан), сюда же попадают Тунис, Куба, Туркменистан, Узбекистан, Казахстан, Белорусия; менее типичными представителями этого сгущения, согласно карте данных, являются Сирия, Алжир, Марокко, Малайзия (эти страны сдвинуты в сторону большего индекса влияния) и Сингапур (сдвинут по градиенту качества жизни).

G17) *Азиатские тоталитарии*: Иран, КНДР, Вьетнам, Саудовская Аравия, Йемен, Ливия, Пакистан, а также Египет. Судя по карте данных, сюда же примыкает Китай, однако от этой группы стран он отличается гораздо большими индексами влияния и государственности, то есть скорее выпадает из этого сгущения (здесь играет свою роль эффект проекции на «край» карты).

I14) *Сгущение азиатских демократий*: Тайланд, Индонезия, Венесуэла, Филипины, а также здесь оказываются Колумбия и Перу; этот кластер отличается относительно высоким уровнем индекса угроз. «Демократичность» этого сгущения надо понимать в относительном смысле (эти страны расположены выше по рейтингу демократичности по сравнению с другими азиатскими странами). К этому сгущению примыкает Турция (но сдвинута в сторону меньших угроз, более высокой государственности и влияния). Индия также оказывается в этом сгущении на карте данных, однако Индия заметно отличается от остальных стран высоким индексом влияния и демократичности.

C02) *мелкие страны Океании* (Тувалу, Палау, Микронезия, Кирибати, Науру);

J01) *Карибский кластер*: Доминика, Гренада, Сант Винсент Гренадины, Санта Лючия, Антигуа и Барбуда, Сант Киттс Невис, а также Белиз и Сейшиллы

E09) «*демократический» Африканский кластер*: Мозамбик, Буркина Фасо, Камерун, Танзания, Бенин, Гамбия, Берег Слоновой Кости, Сенегал, Намибия, Кения; к этому кластеру примыкает *сгущение демократических африканских стран с высоким индексом угроз* (B09): Эфиопия, Того, Сьера Леоне, Бурунди, ЦАР, Чад; также здесь оказываются Гаити и Афганистан. «Демократичность» этого сгущения надо понимать в относительном смысле (эти страны расположены выше по рейтингу демократичности по сравнению с другими африканскими странами).

B12) *«тоталитарный» Африканский кластер*: Заир, Джибути, Мавритания, Сомали, Эритреа, Судан, Ангола, менее типичны Экваториальная Гвинея, Конго, Руанда, Гвинея, Уганда; здесь же оказались Непал, Лаос, Таджикистан и Грузия.

Как следует из карты данных, многие страны не могут быть уверенно отнесены к тому или иному сгущению. Например, Ирак (E13), занимает промежуточное положение между сгущениями “азиатских тоталитарий”, “азиатскими демократиями” и «тоталитарным африканским кластером».

Другие страны обладают уникальными сочетаниями индексов и расположены на достаточном расстоянии от всех остальных стран. К ним относится Россия (сильное влияние и государственность при среднем индексе демократии и уровне жизни), США (абсолютный лидер международного влияния и государственности, при развитой демократии и уровне жизни), Южная Корея (влиятельная держава с развитой государственностью и демократией, но при среднем уровне жизни), Япония (по профилю индексов – это ближайший сосед США, но индексы качества жизни, демократии и государственности меньше США пропорционально примерно на 10%, угрозы выше, влияние меньше).

В выбранной нами метрике (способе измерения расстояния между странами) наиболее близкой к России точкой является Турция, затем Венесуэла, Колумбия и Южная Африка. Однако, несмотря на то, что формально эти страны в пространстве индексов являются ближайшими, Россия достаточно далеко отстоит от каждой из них, существенно опережая каждую из этих стран по рейтингу влияния и государственности. В табл.1 приведены «ближайшие соседи России» при других выборах метрики.

Страны постсоветского пространства распадаются на несколько групп. Как уже было отмечено, бывшие прибалтийские республики и Украина примыкают к сгущению «типичных серых стран». Молдавия и Армения (G04) образуют небольшое сгущение с некоторыми странами бывшей Югославии (Боснией и Герцеговиной, Македонией и Сербией). Белорусия, Казахстан, Узбекистан и Туркменистан попадают в кластер «типичных серых тоталитарий». Азейбарджан (E12) примыкает к сгущениям «африканских и азиатских демократий». Таджикистан, Грузия и Кыргызстан находятся близко к сгущению «африканских тоталитарий».

Пять цветных информационных раскрасок дают представление о поведении того или иного индекса на карте данных. В отличие от географической карты, раскрашенной значением того или иного индекса (см. Приложение к этой работе или веб-сайт **http://atlas.savvy.ru**), эти раскраски меняются вдоль карты непрерывно (это является основным свойством карты данных, так как похожие объекты имеют тенденцию располагаться рядом) и позволяют проанализировать тенденции изменения того или иного индекса в тех или иных «сгущениях» на карте данных. Например, из анализа раскрасок становится понятно, что индекс качества жизни образует примерно горизонтальный градиент цвета слева направо, индекс влияния меняется примерно по диагонали от верхнего левого до нижнего правого угла, а индекс демократии меняется нетривиально, образуя локальный минимум в центре карты (сгущение «типичных тоталитарий») и абсолютный минимум в левом нижнем углу (захватывая сгущения африканских и азиатских тоталитарий)

2. Географические и математические карты

Карта данных представляет из себя «подложку» для визуализации таблицы данных. Идея информационных раскрасок, накладываемых на карту данных, позволяет использовать хорошо разработанные методы и приемы ГИС-технологий (ГИС – геоинформационные системы).

В нашем примере, существует также и реальная географическая карта, на которую мы можем отобразить результаты анализа многомерных данных. Покажем, как на географической карте можно визуализировать результаты анализа главных компонент, а также связать спектральной раскраской карту данных и реальную географическую карту.

2.1. Анализ главных компонент и его визуализация на географической карте

Мы можем считать, что главные компоненты представляют из себя новый базис пространства индексов. Преимущество использования главных компонент в том, что они представляют из себя некоррелированные факторы, причем упорядоченные по силе «объяснения» различий между странами.

Веса индексов для каждой из главных компонент, рассчитанных для таблицы индексов Политического Атласа Современности приведены на рис.2 (внизу, справа). С помощью красного и зеленого цвета показан знак вклада каждого из индексов. Положительный вклад означает, что страны с большим значением соответствующего индекса будут иметь большое значение проекции на соответствующую компоненту и показаны на карте красным цветом, отрицательный вклад даст отрицательную проекцию для соответствующих стран и они будут раскрашены зеленым цветом. Серый цвет страны, означает, что ее индексы сбалансированы относительно тех индексов и с теми весами, которые образуют главную компоненту. При этом надо учитывать, что в нашем анализе значение индекса выше среднего вносит положительный вклад в проекцию, а ниже – отрицательный.

Приведем пример – первая главная компонента, интерпретация которой кажется нам достаточно очевидной (и сходна с приведенной в [1] интерпретацией «противодействие внешним и внутренним угрозам»). В эту компоненту с большими положительными весами входят качество жизни и государственность, и с большим отрицательным – индекс угроз. Существенно меньший положительный вклад дают индексы демократии и влияния. Это означает, что «благополучная страна», с крепким государством и хорошим уровнем жизни, будет обладать большой положительной проекцией на первую компоненту. При этом тип власти («демократия» или «тоталитария») и влиятельность страны на международной арене не играют большой роли. Наоборот, «страна-неудачник» с индексами государственности и качества жизни ниже среднего, и, еще хуже, с высокими значениями индекса угроз, будет обладать отрицательной проекцией на первую главную компоненту. Таким образом, основное различие между странами в современном мире идет вдоль оси «благополучия». Эта ситуация изображена на рис.2 (вверху, слева). Красные страны являются благополучными, причем интенсивность красного пропорциональна «благополучию» страны, то есть рейтингу государственности и качества жизни и обратно пропорциональна рейтингу угроз.

Из рис.2 мы видим, что различие по этой оси в значительной степени связано с географическим положением – в общем, мы имеем «неблагополучную» Африку и часть Азии. Южная Америка, скорее, «благополучна», за исключением Боливии и Гайаны. Лидером благополучности является Швейцария

(значение проекции 4.0). Западно- и североевропейские страны, США, Канада, Австралия имеют примерно одинаковую проекцию на «благополучную» первую компоненту (3.0-3.5), Россия обладает «благополучием», равным 1.2, Япония – 2.9, Китай – 0.5, Индия – 0.35. Среди наиболее неблагополучных стран – страны центральной и западной Африки (Чад, Мавритания, ЦАР, Сомали и другие), а также Афганистан, Таджикистан, Кыргызстан («неблагополучие», которое здесь во многом связано с обеспечением питьевой водой, лежит в интервале -2.9 – -2.5). На Кавказе неблагополучна Грузия (-2.1), в Южной Азии – Лаос (-2.1).

Ряд стран (например, в северной Африке, Иран, Ботсвана, Индонезия) обладают «благополучем» близким к нулю. На карте рис.2 (вверху, слева) они изображены серым цветом.

Менее очевидна интерпретация второй и последующих главных компонент. Во второй компоненте индекс влияния имеет большой отрицательный вклад (-0.81), а – демократичности положительный (0.41). Качество жизни для этой компоненты не играет роли. Обращая знаки, можно сказать, что это компонента «неконтролируемой силы», то есть силы, не сдерживаемой демократическими институтами. Неудивительно, что лидерами вдоль этой компоненты (ярко-зеленые страны на рис.2, вверху, справа) являются такие страны как Китай (проекция отрицательна и равна -2.9), КНДР (-2.1), Саудовская Аравия (-2.2). На обратном полюсе на этой компоненте находятся демократические страны, не имеющие большого влияния в мире: многие страны Океании и микрогосударства Европы (Лихтенштейн, Андорра, Монако) имеют проекцию порядка +2, страны Прибалтики, Молдова, Армения, Монголия, Исландия имеют проекцию примерно +1. США и Россия обладают примерно равной проекцией на эту компоненту (примерно -1.8), Япония - -1.6. Осложняют интерпретацию также существенный отрицательный по знаку вклад индекса угроз и государственности. Таким образом, некоторые страны Африки и Ближнего Востока также получают сравнительно большую, примерно -1.5, отрицательную проекцию (Судан, Эфиопия, Йемен, Ирак). Смысл этого эффекта неясен.

Недостаточно прозрачна интерпретация и третьей компоненты. Здесь отрицательным весом обладают индексы демократичности и угроз. Им противостоит индекс уровня жизни. Таким образом, ярко красными странами оказываются богатые и благополучные тоталитарии (ОАЭ, проекция - +2.2, султанат Бруней и Даруссалам - +2.3, эмират Катар - +2.2), а ярко-зелеными – бедные демократические государства, борющиеся с внешними и внутренними трудностями (Индия - -1.9, Эфиопия - -1.5). Таким образом, большими проекциями в этой компоненте обладают страны, выпадающие из усредненной корреляционной связи между индексами качества жизни и демократии (страны с высоким уровнем жизни имеют тендецию развивать демократические институты) и связи индекса угроз и демократичности (высокие угрозы препятствуют развитию демократии). Условно обозначим смысл этой компоненты как «трудности демократии».

В четвертой компоненте внешнее влияние (отрицательный вклад) противостоит индексу угроз и примерно равным вкладам со стороны индексов государственности и качества жизни (вклад демократичности пренебрежимо мал). Можно согласиться с интерпретацией этого фактора, приведенной в [1] как «влияние во чтобы то ни стало». В контексте нашего анализа можно уточнить: «международное влияние, несмотря на собственные проблемы». Страны с большой отрицательной проекцией обладают влиянием на мировой арене, непропорционально большим, по сравнению с уровнем внутренних проблем, которые весьма серьезны, (например, Сербия, проекция – -1.34, Украина – -1.17). Интересно, что на этой компоненте среди «зеленых» стран – многие страны бывшего СССР (Украина, Молдавия, Армения, Беларусь, Казахстан, Латвия и Литва) и почти все страны бывшего социалистического лагеря (Сербия, Босния и Герцеговина, Болгария, Хорватия, Чехия, Словакия, Польша, Румыния, Македония, Венгрия). Россия на этой компоненте входит в «серую» зону, то есть ее международное влияние объяснимо и пропорционально «объективной» силе государства. На противоположном полюсе либо благополучные страны, степень международного влияния которых непропорционально мала по каким-то причинам (Люксембург – +1.3, Исландия – +0.9) или бедственные страны, не имеющие никакого влияния (Мьянмар - +1.0, Зимбабве - +0.9, Филиппины – +0.9). Интересно, что лидеры мирового влияния – Япония и США обладают заметными положительными проекциями (+0.93 и +0.62), то есть согласно этой компоненте их международное влияние «недостаточно» велико по сравнению с усредненной пропорцией. Это, по видимому, указывает на нелинейность «функции влияния» от индекса государственности и угроз (надо иметь в виду, что мы измеряем индекс влияния в логарифмической шкале).

Пятая главная компонента связана с балансом «государственность-качество жизни». Красными на рис.2 (третья строка, слева) оказываеются те страны, где этот баланс решается в сторону уровня жизни, зелеными – страны, развивающие в первую очередь государственность. На одном полюсе этой оси оказываются страны с очень высоким уровнем жизни (например, Люксембург, проекция +1.5, Норвегия, проекция +0.8) или чрезвычайно слабые государства (например, Киргизстан - +1.26, Грузия - +0.92). На другом полюсе – крепкие государства со средним уровнем жизни (например, Ямайка – -0.9, Бразилия – -0.8). По этой компоненте разделены Россия (проекция -0.3) и США (+0.4). Примеры сбалансированных в этом отношении, «серых» стран: Саудовская Аравия, Индия, Казахстан, Белоруссия, КНДР.

2.2. Раскраска географической карты на основе «карты данных»

Как мы видим, главные компоненты обладают различным значением, интерпретация которого далеко не всегда очевидна. Главные компоненты, рассчитанные в нашем анализе данных, несколько отличаются от приведенных в [1], по причине альтернативного выбора метрики (см. раздел 3.1). Это подчеркивает зависимость значения и интерпретации главных компонент от выбора метрики пространства. Отсутствие предобработки индекса влияния в [1] делает его статистически менее значимым (меньший вклад в дисперсию), поэтому в [1] индекс влияния значимо влияет только на 4ую компоненту, в то время как в нашем анализе индекс влияния значительно определяет поведение уже второй главной компоненты. Некоторые компоненты являются более «стабильными» к изменению метрики, хотя и могут изменить порядок и величину «объясненной» дисперсии. Так, легко можно идентифицировать компоненты 1 и 4 в нашем анализе и те же компоненты в анализе [1]. Пятая главная компонента в нашем анализе (баланс «государственность-качество жизни») может быть ассоциирована с компонентой 3 в анализе, проведенном в [1].

Возможно, правильнее интерпретировать последовательность нескольких первых компонент: первая разделяет страны по усредненному «благополучию», вторая вносит в это разделение аспект, связанный с демократичностью и влиянием, третья описывает нюансы, связанные с необычным, «странным» богатством некоторых стран (как Монако, ОАЭ, Кувейт) или наоборот, бедностью других (Индия, Эфиопия) и т.д. Главные компоненты не являются статистически независимыми в строгом смысле слова (см. [9]), поэтому следует осторожно относится к интерпретации, например, отдельно взятой четвертой главной компоненты.

Используя возможности компьютерной графики, можно наложить несколько первых компонент и визуализировать их одновременно. Так, например, в Приложении к этой работе приведена иллюстрация, на которой первая главная компонента визуализирована в красном канале цвета, вторая – в зеленом, и третья – в синем. Таким образом, страны, обладающие похожими проекциями на первые три главные компоненты, будут иметь похожие цвета. Большая положительная проекция соответствует максимальной яркости соответствующего цвета, большая отрицательная соответствует отсутствию цвета в канале. Например, Китай и Саудовская получают одинаковые фиолетовые цвета. Положительная проекция на компоненту 1 (это в целом «благополучные» страны) дает красный цвет, отрицательная проекция на компоненту 2 (это влиятельные тоталитарии) убирает на ноль зеленый канал, положительная проекция на компоненту 3 (несмотря на сравнительно высокое качество жизни, эти страны не развивают демократии) дает синий цвет. Смешение красного и синего без зеленого дает фиолетовый цвет этих двух стран.

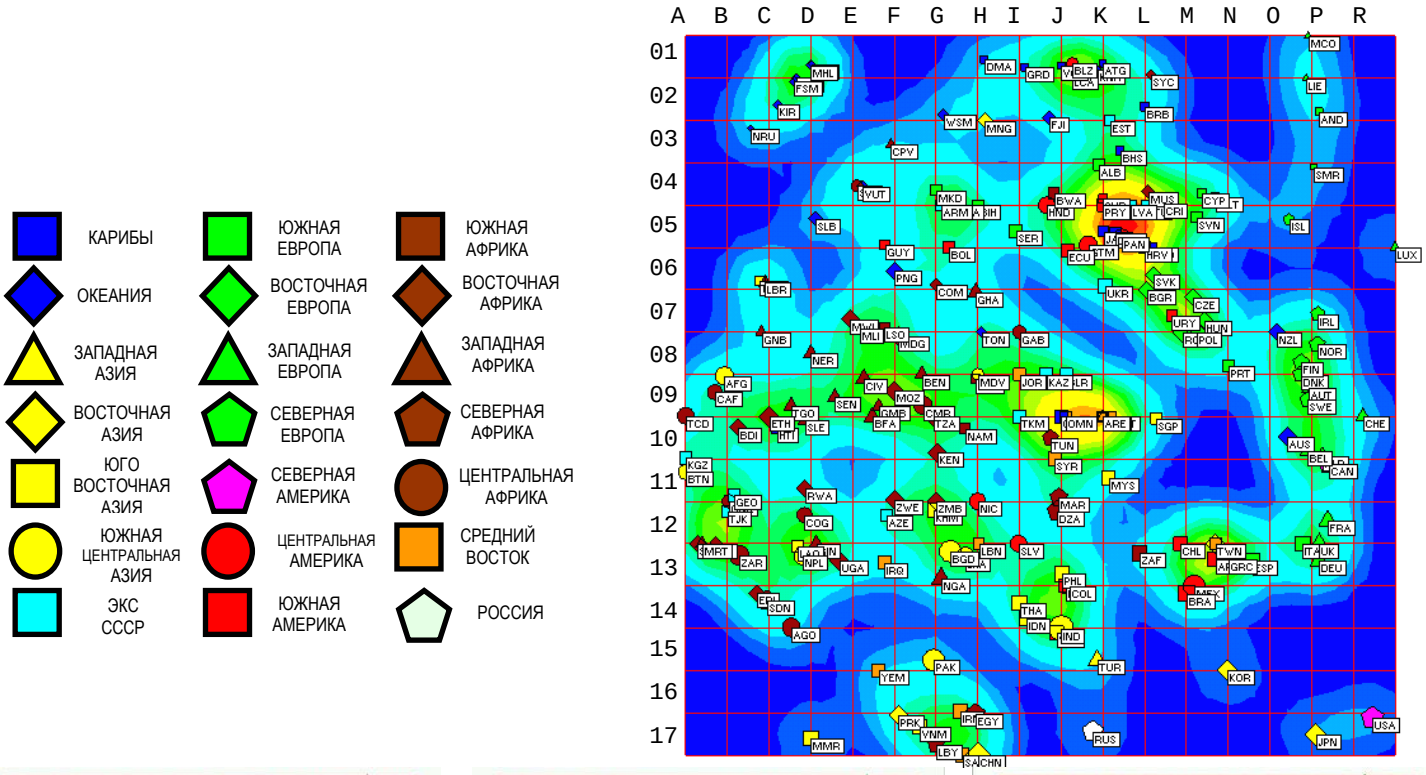
Аналогичный прием позволяет визуализировать одновременно три выбранных индекса и их соотношение. Так, например, в Приложении приведена иллюстрация, на которой нанесена раскраска по индексам демократичности в красном канале, международного влияния в зеленом и качества жизни в синем. Так, например, светло-салатовый цвет США на этой раскраске говорит об одновременно высоком значении всех трех индексов, но с преобладанием индекса влияния и качества жизни (салатовый оттенок). Темно-красный цвет Монголии говорит о том, что эта страна может «похвастаться» только некоторым развитием демократии, а влияние и качество жизни имеет значительно ниже среднего. Африка в целом темнее (меньшие в среднем значения индексов), Европа - светлее. Страны со сходными значениями этих трех индексов обладают похожими цветами.

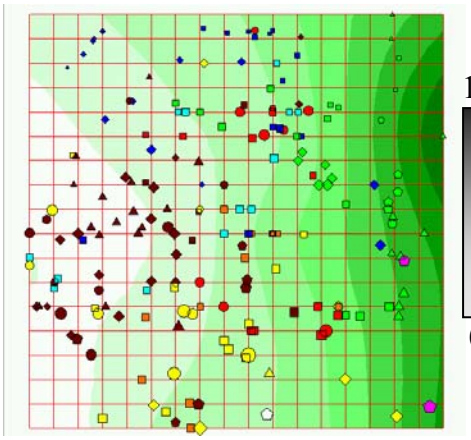
Предложим еще один прием, который позволяет перенести результаты картографирования данных на географическую карту (см. рис.3). На двумерную карту данных наносится двумерный спектр таким

образом, чтобы точки, расположенные рядом на карте оказались бы в областях схожего цвета. Полученные цвета используются для раскраски географической карты. Преимущество такого способа раскраски от описанного двумя параграфами выше заключается в том, что в нем, как и в построении карты, потенциально участвуют информация из всех пяти, а не только первых трех компонент. Однако, прямая интерпретация цвета невозможна, можно лишь сказать какие страны больше или меньше похожи друг на друга. Аналогичная технология была применена группой Кохонена в 1992 году при анализе 39 показателей качества жизни для 126 стран мира (см. ссылку в Интернете **http://www.cis.hut.fi/research/som-research/worldmap.html**).

2.3. Онлайновая ГИС-система с результатами анализа

Для повышения наглядности иллюстраций и интерактивности нами разработан веб-сайт **http://atlas.savvy.ru**, на котором можно запустить простую ГИС-систему, позволяющую визуализировать результаты многомерного анализа данных, как на карте данных, так и на географической карте. Эта система позволяет в интерактивном режиме получить значения индексов для любой страны, ее проекции на главные компоненты, а также с помощью щелчка компьютерной мыши открыть соответсвующую статью из свободной онлайновой энциклопедии «Википедия». Разработанная нами ГИС-система может быть также установлена на локальном компьютере пользователя в виде веб-приложения и базы данных с информацией для раскрасок.

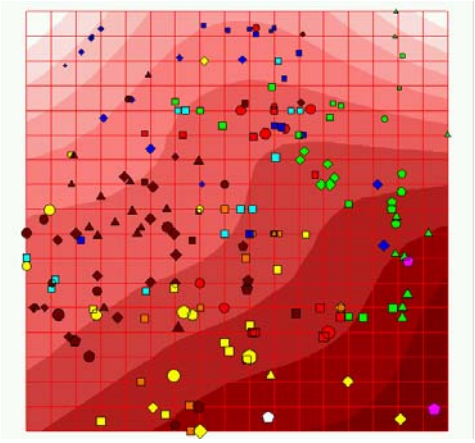




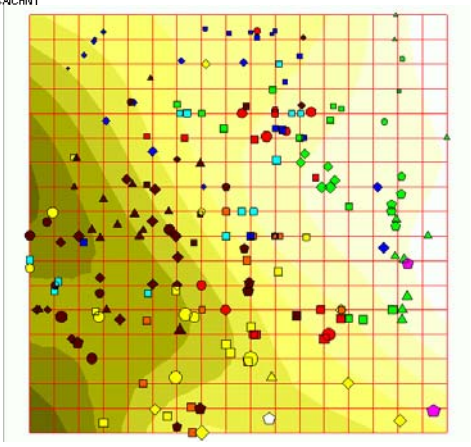
10

0

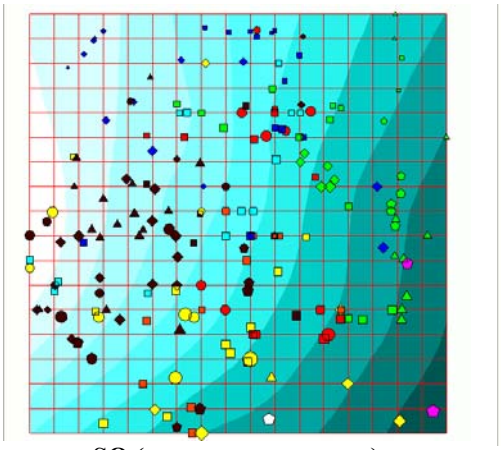
QL (качество жизни)



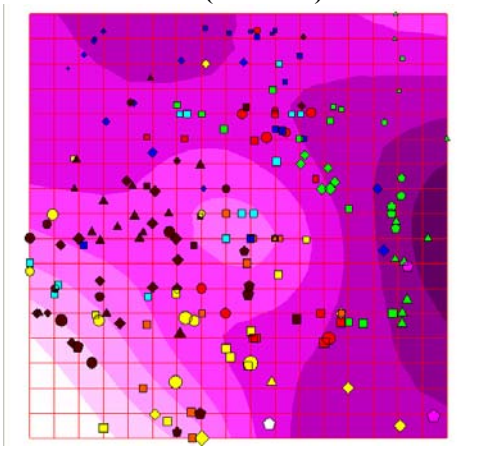
PW (влияние)



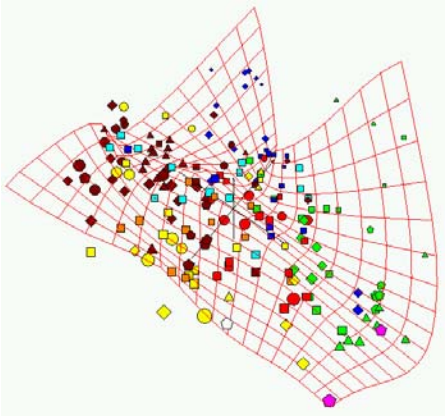
MN (угрозы)



SO (государственность)

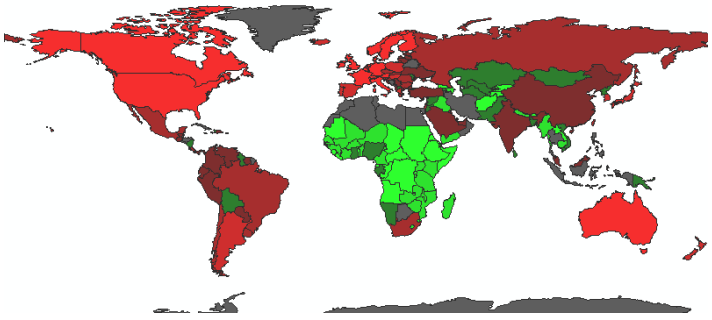


DM (демократия)

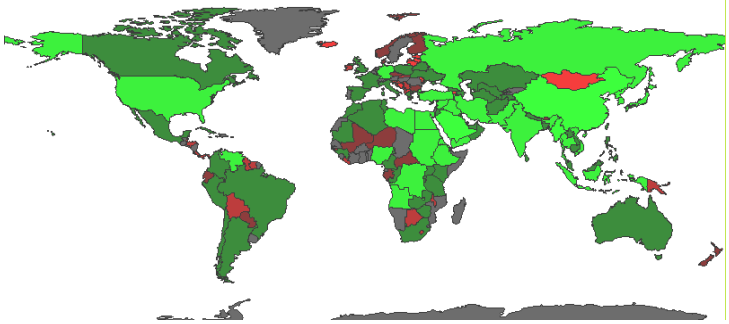


вид карты в пространстве главных компонент

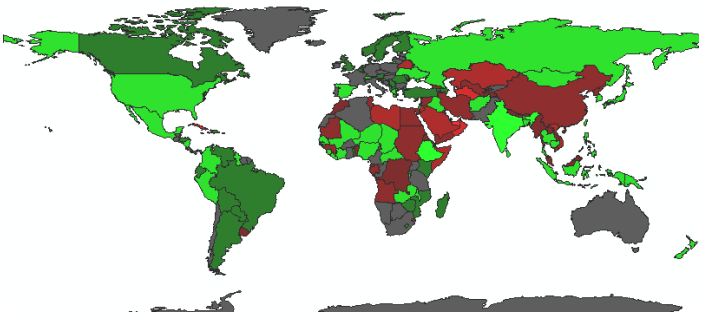
Рис.1. Карта данных индексов Политического Атласа Современности и атлас информационных раскрасок. Цветом и формой отмечены различные группы стран. Размер точки пропорционален логарифму населения страны. Близкие точки соответствуют странам с похожими значениями индексов. На раскраске вверху справа цветом изображена плотность точек на карте, что позволяет выявить кластеры или «сгущения» в распределении данных. На остальных раскрасках показано изменение значения того или иного индекса.
Интерактивная версия этой иллюстрации доступна на http://atlas.savvy.ru .



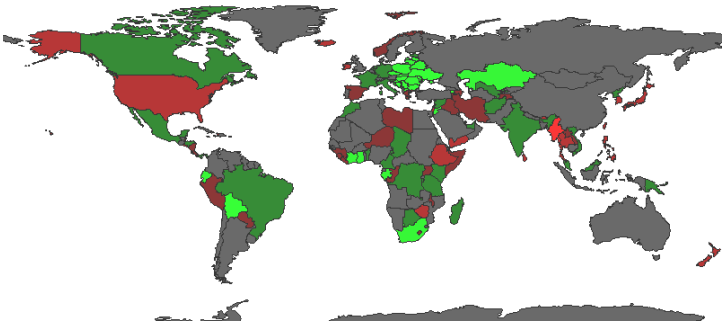
главная компонента 1



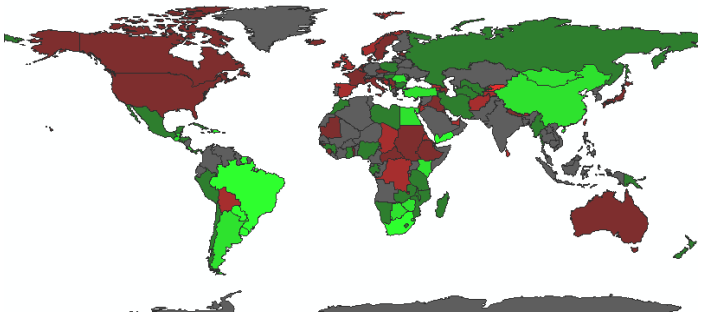
Главная компонента 2



главная компонента 3



Главная компонента 4



главная компонента 5

веса индексов в главных компонентах

	PC1	PC2	PC3	PC4	PC5
Объясненная дисперсия	0.59	0.21	0.12	0.04	0.04
Кумулятивная объясненная дисперсия	0.59	0.80	0.92	0.96	1.00
QL (качество жизни)	0.53	0.11	0.24	0.45	0.67
PW (межд.влияние)	0.28	-0.81	-0.25	-0.37	0.25
MN (угрозы)	-0.48	-0.31	-0.43	0.69	0.12
SO (государственность)	0.52	-0.26	0.10	0.43	-0.69
DM (демократия)	0.38	0.41	-0.83	-0.07	-0.03

Рис. 2. Анализ главных компонент данных рейтингов “Политического Атласа Современности” и визуализация проекций (нагрузок) стран на главные компоненты. Красный цвет отображает положительные проекции, зеленый – отрицательные, серый – близкие к нулю, интенсивность цвета пропорциональна абсолютному значению проекции.

Попытка интерпретации смысла главных компонент проделана в разделе 2.1. Ниже мы даем лишь резюме этой интерпретации

Главная компонента 1 : «благополучие» страны; красные страны «благополучны», зеленые - нет

Главная компонента 2 : «неконтролируемая сила», зеленые страны – в них международное влияние нескомпенсировано развитием демократических институтов, а также бедствующие страны.

Главная компонента 3 : «трудности демократии», красные страны – богатые тоталитарные режимы, красные – демократические режимы, испытывающие трудности

Главная компонента 4 : «международное влияние, несмотря на собственные проблемы» - это девиз «зеленых» стран, в то время как «красные страны» имеют нереализованный потенциал влияния

Главная компонента 5 : «ось государственность-качество жизни», в зеленых странах этот баланс решен в пользу государственности, в красных – в пользу качества жизни.

Интерактивная версия этой иллюстрации доступна на http://atlas.savvy.ru

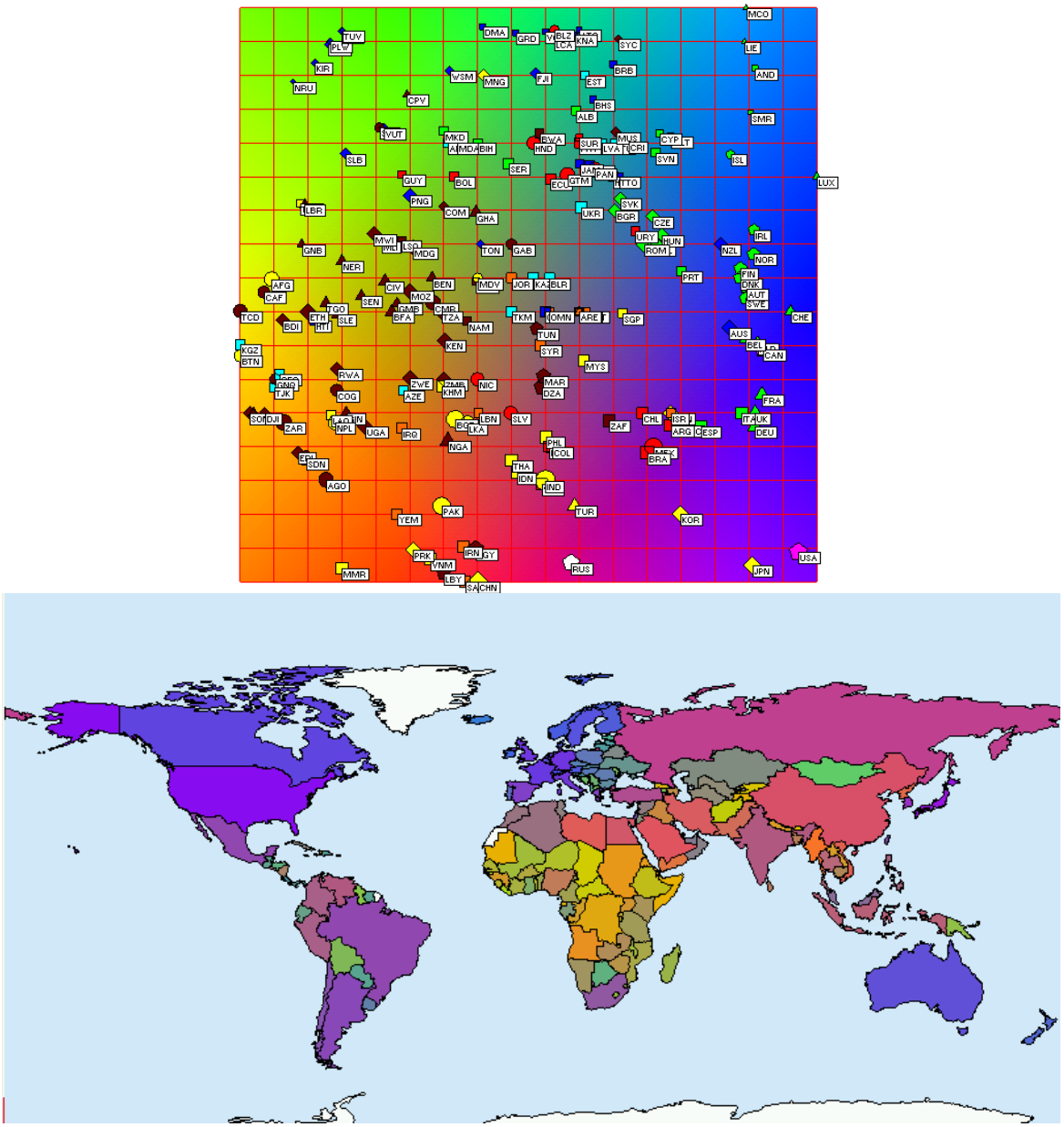


Рис. 3. Проекция карты данных на географическую карту. На двумерное распределение точек на карте данных нанесен «двумерный» спектр так, чтобы близкие точки попадали в области с похожим цветом. Полученные цвета используются для раскраски географической карты. Таким образом, страны с близкими цветами имеют похожие комбинации индексов.

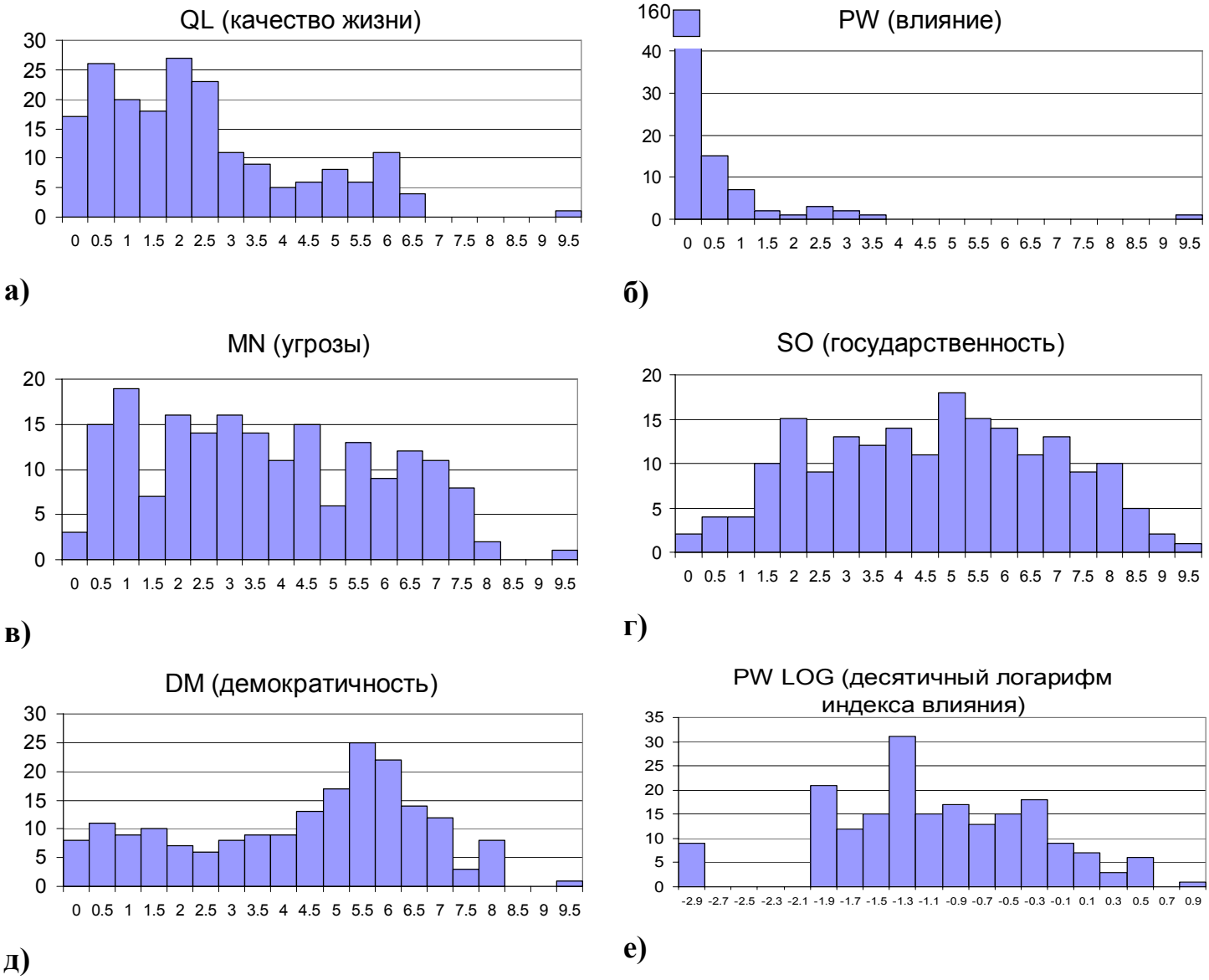


Рис. 4. Гистограммы распределений индексов. Индексы качества жизни (а) и демократии (д) распределены бимодальным образом. Таким образом проявляются основные дуальности в современном мире («богатый-бедный», «демократия-тоталитаризм»). Распределение индекса влияния (б) близко к геометрическому («распределение с тяжелым хвостом») так как сравнительно небольшая группа стран (большая «восьмерка» и дополнительно Китай и Индия, за счет большого населения) обладают высокими значениями этого индекса. Индекс угроз распределен примерно равномерно по странам (в), а государственности – по распределению, близкому к нормальному, однако с заметным положительным вкладом высших четных моментов. (е) Условия применимости метода главных компонент можно улучшить, если измерять индекс влияния в логарифмической шкале (нулевое значение индекса заменено на 10^{-3}).

3. Методологические аспекты анализа рейтингов

3.1. Выбор метрики в пространстве индексов государств

Выбор метрики, то есть способа измерения расстояния между точками-государствами в пространстве индексов, является принципиальным вопросом, который необходимо решить перед тем, как применять любые методы многомерного анализа. Проблема заключается в следующем: как для двух выбранных стрн аггрегировать различия в пяти индексах-координат пространства в одно значение «расстояния», которое бы адекватно отражало различия между этими странами? Метрика определяет геометрические свойста пространства и любые выводы, полученные с помощью таких методов анализа как метод главных компонент или кластерный анализ, в большой степени определяются этим выбором.

Ни в коем случае не умаляя результатов исследования [1], стоит все-таки отметить, что многие выводы [1] зависят от конкретного выбора параметров в применении методов анализа, в частности, от выбора метрики пространства индексов. В работе [1] выбран один из возможных вариантов (который, надо отметить, недостаточно подробно запротоколирован) и не произведен анализ того, как полученные выводы могли бы измениться при другом выборе метрики. В этом разделе мы рассматриваем этот вопрос и обосновываем способ измерения расстояний, несколько отличный от сделанного в [1]. Не настаивая на оптимальности нашего выбора, мы верим в то, что анализ проведенный с различных точек зрения, помогает отделить действительно значимые выводы от «фокусов статистики», то есть феноменов, сильно зависящих от выбора конкретных параметров метода.

Приведем конкретный пример. Рассмотрим несколько выборов метрики: 1) евклидова, примененная к исходным индексам таблицы (под *значениями* индексов мы понимаем численный «балл» того или иного индекса, см.Приложение к работе [1], не стоит путать значение индекса и *рейтинг* государства, то есть порядковый номер страны после упорядочения в порядке убывания того или иного индекса); 2) L1 (то есть сумма абсолютных значений разностей индексов), примененная к исходным значениям индексов; 3) евклидова примененная к z-значениям (z-значение получается путем вычитания среднего значения переменной и деления его на стандартное отклонение) и после логарифмического преобразования индекса влияния (мы использовали эту метрику в нашем анализе); 4) то же, что и 3, но L1 вместо эвклидова расстояния; 5) метрика L1, примененная к значениям рейтингов (то есть сумма различий в положении страны по разным рейтингам); 6) единица минус коэффициент корреляции между индексами (индекс влияния в логарифмической шкале), в этой метрике роль играет не абсолютное значение индексов, а их «профиль», то есть изменения индексов один относительно другого. Эти метрики используются для того, чтобы рассмотреть «ближайших соседей» России в пространстве индексов. Результаты приведены в таблице 1.

Стандартные методы факторного анализа, такие как метод главных компонент, подразумевают евклидову метрику пространства (более общо, квадратичную метрику). Существует свобода в выборе коэффициентов квадратичной метрики, а также свобода в выборе способа предварительного преобразования исходных значений координат. Следует упомянуть о существовании методов для обучаемого выбора таких преобразований (в частности, для выбора весов во взвешенной евклидовой метрике). Например, в [8] и [2] описан метод , при котором составляются пары «похожих» точек (стран в нашем случае) и веса метрики оптимизируются, чтобы похожие точки оказались по возможности на близких расстояниях в результирующем евклидовом пространстве.

В этой работе для определения метрики мы опираемся на стандартный анализ гистограмм распределений значений признаков с тем, чтобы выбором метрики обеспечить наилучшие условия применимости таких методов как анализ главных компонент. На рис.4 представлены гистограммы исходных значений пяти индексов. Интересно отметить, что индексы качества жизни и демократии имеют бимодальный характер распределения, индекс государственности распределен примерно по гауссиане (однако все же, с визуально заметным вкладом высших, начиная с четвертого, моментов), характер распределения индекса угроз можно охарактеризовать как близкий к равномерному. Индекс международного влияния имеет выраженный характер геометрическим распределения, имеющего «длинный хвост». Это отражает тот факт, что относительно малое число стран имеют реальную силу на международной арене. Подобный характер распределения одного из признаков в таблице данных вносит существенную нелинейность «не по существу» в геометрическую структуру облака данных, делая применение, например, метода главных компонент, менее обоснованным. Стандартным приемом в данном случае является логарифмическое преобразование данного признака, с тем, чтобы сделать более его гистограмму распределения более близкой к нормальной (см рис. 4e).

Из рис.4 и простого расчета (данные не приведены) следует, что стандартные отклонения признаков сравнимы для признаков QL, MN, SO и DM (равны примерно 2.0). Для индекса PW (который измеряется в логарифмической шкале) стандартное отклонение равно 0.8. Это означает, что без преобразования к z-значениям (центрирование и нормировка на стандартное отклонение), индекс PW вносил бы заметно меньший вклад в расстояния между государствами, по сравнению с остальными признаками. Для выравнивания значимости признаков мы применили стандартное преобразование к z-значениям.

3.2. Взвешивание точек данных

Эффективным аналитическим инструментом в многомерном анализе может служить приписывание весов каждой точке данных. Анализ главных компонент, кластерный анализ и технология картографирования данных естественным образом обобщаются на случай взвешенных точек (см., например, [2,3]). Приведем пример использования взвешивания для анализа данных Атласа.

Одно из возможных критических замечаний в адрес анализа данных, представленных в журнале «Полис» и в этой статье - это тот факт, что все страны вносят одинаковый вклад в расчет главных компонент. Таким образом, при учете мировых тенденций, одинаковую роль играет миллиардный Китай и крошечная Тувалу с 12ю тысячами жителей, что, конечно же, трудно оправдать с точки зрения методологии. Выход из этой ситуации – приписать точке-стране вес, равный числу ее жителей. Таким образом, равный вклад в расчет главных компонент будет вносить не страна, а каждый ее житель. Результат такого анализа приведен в материалах Приложения. В частности, очевидным становится существование в мире двух полюсов силы – «демократического», во главе с США и странами большой восьмерки, и «тоталитарного», во главе с Китаем и мощными странами Азии, Ближнего и Среднего Востока (КНДР, Пакистан, Иран, Саудовская Аравия). Ясно, что по числу человек, вовлеченных в эти два блока, они приблизительно равносильны, однако значительно отличаются по отношению к росту благосостояния населения, составляющего тот и другой блок (индекс качества жизни QL не вносит почти никакого вклада во вторую компоненту, которая «притягивается» Китаем).

Анализ главных компонент в случае взвешивания точек по количеству населения не меняет смысла первой, «благополучной» компоненты, а также почти не меняет знаков вкладов индексов в других компонентах. Однако, абсолютное значение вклада индекса может существенно измениться, что может привести с изменению «смысла» компоненты. Так, например, .во второй компоненте акцент смещается с «влияния» на «демократию», сводится к нулю роль индекса угроз. Таким образом, вторая компонента вместо интерпретации «неуправляемая сила» получает, скорее, интерпретацию «организованная тоталитария». Несколько изменяется смысл третьей компоненты. Смысл четвертой компоненты остается примерно тем же (однако, с поправкой, «влияние несмотря ни на что, но не затрагивая государственности»), а в пятой государственность противостоит не столько качеству жизни, сколько индексу влияния.

Заключение

Также, было бы небезынтересно проанализировать исходные данные по 60ти индикаторам для подробного рассмотрения взаимосвязей в 60-мерном пространстве. По мере повышения размерности пространства человеческая интуиция работает хуже, а современные методы факторного и кластерного анализа становятся более актуальными.

В заключение, авторы выражают благодарность компании «Новые коммуникационные системы» (г. Екатеринбург) за поддержку проведенного исследования.

Литература

1. Мельвиль А.Ю., Ильин М.В., Мелешкина Е.Ю., Миронюк М.Г., Полунин Ю.А., Тимофеев И.Н.. Опыт классификации стран. 2006, *Полис* **5**, с. ??. Статья доступна онлайн: http://www.polis.ru/ . См. также сайт проекта http://worldpolities.org/ .
2. Зиновьев А.Ю. Визуализация многомерных данных. Красноярск, изд-во КГТУ, 2000. Монография доступна онлайн: http://www.ihes.fr/~zinovyev/papers/book/ZinovyevA_Visualization_of_multidimensional_data_2000.pdf .
3. Gorban A., Zinovyev A. Elastic Principal Graphs and Manifolds and their Practical Applications. 2005. *Computing* **75**, 359 - 379.
4. Горбань А.Н., Зиновьев А.Ю., Питенко А.А. Визуализация данных методом упругих карт. *Информационные технологии* **6**, 2000. с. 26-35.
5. Gorban A.N., Zinovyev A.Yu. Visualization of data by method of elastic maps and its applications in genomics, economics and sociology. *Institut des Hautes Etudes Scientiques preprint*, France. 2001. M/01/36. Препринт доступен онлайн: http://www.ihes.fr/~zinovyev/papers/M01-36.pdf .
6. Зиновьев А.Ю., Питенко А.А., Попова Т.Г. Практическое применение метода упругих карт. *Нейрокомпьютеры* **4**, 2002. с. 31-39.
7. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
8. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Статистическое оценивание зависимостей. – М.: Финансы и статистика, 1985. – 484 с.
9. Oja E., Hyvaryinen A., Karhunen J. Independent Component Analysis. 2001. Wiley Interscience, 504 pages.

Таблица 1. Таблица 10 «ближайших соседей» России в пространстве индексов при различном выборе метрики

N	Евклидова Z-values Log PW		L1 Z-values Log PW		Евклидова		L1		L1 рейтинги		1-r Log PW	
	рст	страна	рст	страна	рст	Страна	рст	страна	рст	страна	рст	страна
1	0.54	Турция	1.00	Турция	1.58	Турция	1.49	Турция	54	Турция	0.001	Турция
2	0.94	Венесуэла	1.69	Венесуэла	2.33	Венесуэла	2.71	Венесуэла	80	Венесуэла	0.015	Марокко
3	1.24	Колумбия	1.99	Колумбия	2.58	Перу	2.78	Перу	93	Колумбия	0.018	Венесуэла
4	1.26	Бразилия	2.08	Перу	2.69	Колумбия	2.95	Колумбия	107	Перу	0.018	Гватемала
5	1.31	Индонезия	2.24	Южная Африка	2.85	Доминиканская Респ	3.53	Южная Африка	124	Южная Африка	0.020	Габон
6	1.33	Африка	2.42	Бразилия	2.89	Парагвай	3.74	Доминиканская Респ	125	Египет	0.020	Доминиканская Респ
7	1.35	Перу	2.61	Египет	2.97	Гватемала	4.11	Гватемала	128	Республика Корея	0.025	Парагвай
8	1.41	Республика Корея	2.66	Индонезия	2.97	Южная Африка	4.30	Парагвай	130	Китай	0.025	Ямайка
9	1.42	Мексика	2.67	Республика Корея	3.02	Бразилия	4.46	Чили	133	Иран	0.028	Перу
10	1.54	Чили	2.70	Мексика	3.03	Ямайка	4.56	Ямайка	141	Бразилия	0.031	Южная Африка

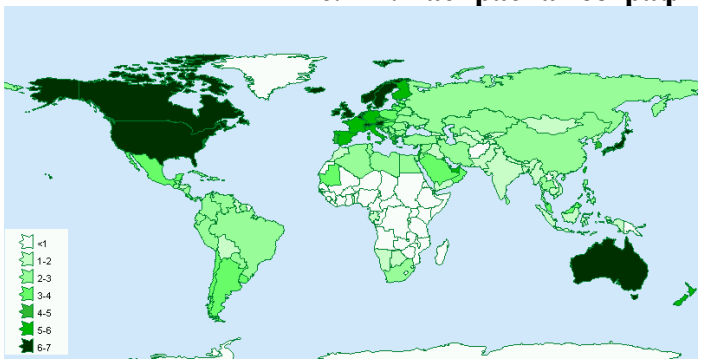
ПРИЛОЖЕНИЯ

Таблица П1. Таблица трехбуквенных кодов ISO и двубуквенных кодов FIPS

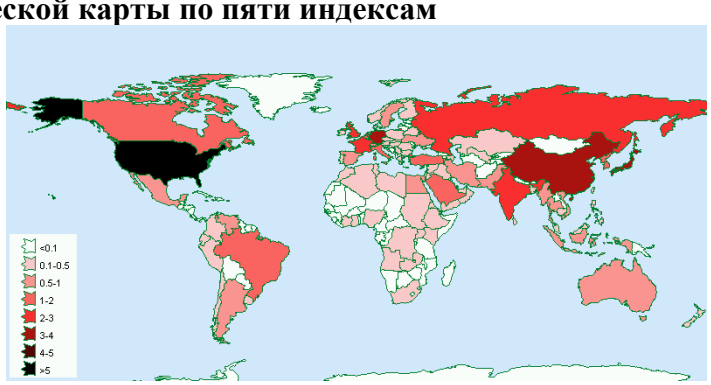
Страна	ISO	FIPS	Страна	ISO	FIPS	Страна	ISO	FIPS
Афганистан	AFG	AF	Гвинея Биссау	GNB	PU	Науру	NRU	NR
Ангола	AGO	AO	Экватор Гвинея	GNQ	EK	Новая Зеландия	NZL	NZ
Албания	ALB	AL	Греция	GRC	GR	Оман	OMN	MU
Андорра	AND	AN	Гренада	GRD	GJ	Пакистан	PAK	PK
ОАЭ	ARE	TC	Гватемала	GTM	GT	Панама	PAN	PM
Аргентина	ARG	AR	Гайана	GUY	GY	Перу	PER	PE
Армения	ARM	AM	Гондурас	HND	HO	Филиппины	PHL	RP
Антигуа и Барбуда	ATG	AC	Хорватия	HRV	HR	Палау	PLW	PS
Австралия	AUS	AS	Гаити	HTI	HA	Папуа Новая Гвинея	PNG	PP
Австрия	AUT	AU	Венгрия	HUN	HU	Польша	POL	PL
Азербайджан	AZE	AJ	Индонезия	IDN	ID	КНДР	PRK	KN
Бурунди	BDI	BY	Индия	IND	IN	Португалия	PRT	PO
Бельгия	BEL	BE	Ирландия	IRL	EI	Парагвай	PRY	PA
Бенин	BEN	BN	Иран	IRN	IR	Катар	QAT	QA
Буркина Фасо	BFA	UV	Ирак	IRQ	IZ	Румыния	ROM	RO
Бангладеш	BGD	BG	Исландия	ISL	IC	Россия	RUS	RS
Болгария	BGR	BU	Израиль	ISR	IS	Руанда	RWA	RW
Бахрейн	BHR	BA	Италия	ITA	IT	Саудовская Аравия	SAU	SA
Багамы	BHS	BF	Ямайка	JAM	JM	Судан	SDN	SU
Босния и Герцеговина	BIH	BK	Иордания	JOR	JO	Сенегал	SEN	SG
Белоруссия	BLR	BO	Япония	JPN	JA	Сербия и Черногория	SER	SR
Белиз	BLZ	BH	Казахстан	KAZ	KZ	Сингапур	SGP	SN
Боливия	BOL	BL	Кения	KEN	KE	Соломоновы о ва	SLB	BP
Бразилия	BRA	BR	Кыргызстан	KGZ	KG	Сьерра Леоне	SLE	SL
Барбадос	BRB	BB	Камбоджа	KHM	CB	Сальвадор	SLV	ES
Бруней Даруссалам	BRN	BX	Кирибати	KIR	KR	Сан Марино	SMR	SM
Бутан	BTN	BT	Сент Киттс и Невис	KNA	SC	Сомали	SOM	SO
Ботсвана	BWA	BC	Республика Корея	KOR	KS	Сан Томе и Принсипи	STP	TP
ЦАР	CAF	CT	Кувейт	KWT	KU	Суринам	SUR	NS
Канада	CAN	CA	Лаос	LAO	LA	Словакия	SVK	LO
Швейцария	CHE	SZ	Ливан	LBN	LE	Словения	SVN	SI
Чили	CHL	CI	Либерия	LBR	LI	Швеция	SWE	SW
Китай	CHN	CH	Ливия	LBY	LY	Свазиленд	SWZ	WZ
Кот д`Ивуар	CIV	IV	Сент Люсия	LCA	ST	Сейшельские о ва	SYC	SE
Камерун	CMR	CM	Лихтенштейн	LIE	LS	Сирия	SYR	SY
Конго	COG	CF	Шри Ланка	LKA	CE	Чад	TCD	CD
Колумбия	COL	CO	Лесото	LSO	LT	Того	TGO	TO
Коморские о ва	COM	CN	Литва	LTU	LH	Таиланд	THA	TH
Кабо Верде	CPV	CV	Люксембург	LUX	LU	Таджикистан	TJK	TI
Коста Рика	CRI	CS	Латвия	LVA	LG	Туркменистан	TKM	TX
Куба	CUB	CU	Марокко	MAR	MO	Тимор Леште	TMP	TT
Кипр	CYP	CY	Монако	MCO	MN	Тонга	TON	TN
Чехия	CZE	EZ	Республика Молдова	MDA	MD	Тринидад и Тобаго	TTO	TD
Германия	DEU	GM	Мадагаскар	MDG	MA	Тунис	TUN	TS
Джибути	DJI	DJ	Мальдивские о-ва	MDV	MV	Турция	TUR	TU
Доминика	DMA	DO	Мексика	MEX	MX	Тувалу	TUV	TV
Дания	DNK	DA	Маршалловы о ва	MHL	RM	Тайвань	TWN	TW

Доминиканская Респ	DOM	DR	БЮР Македония	MKD	MK	Объед Респ Танзания	TZA	TZ
Алжир	DZA	AG	Мали	MLI	ML	Уганда	UGA	UG
Эквадор	ECU	EC	Мальта	MLT	MT	Великобритания	UK	UK
Египет	EGY	EG	Мьянма	MMR	BM	Украина	UKR	UP
Эритрея	ERI	ER	Монголия	MNG	MG	Уругвай	URY	UY
Испания	ESP	SP	Мозамбик	MOZ	MZ	США	USA	US
Эстония	EST	EN	Мавритания	MRT	MR	Узбекистан	UZB	UZ
Эфиопия	ETH	ET	Маврикий	MUS	MP	Сент Винсент и Грена	VCT	VC
Финляндия	FIN	FI	Малави	MWI	MI	Венесуэла	VEN	VE
Фиджи	FJI	FJ	Малайзия	MYS	MY	Вьетнам	VNM	VM
Франция	FRA	FR	Намибия	NAM	WA	Вануату	VUT	NH
Микронезия	FSM	FM	Нигер	NER	NG	Самоа	WSM	WS
Габон	GAB	GB	Нигерия	NGA	NI	Йемен	YEM	YM
Грузия	GEO	GG	Никарагуа	NIC	NU	Южная Африка	ZAF	SF
Гана	GHA	GH	Нидерланды	NLD	NL	Дем Респ Конго	ZAR	CG
Гвинея	GIN	GV	Норвегия	NOR	NO	Замбия	ZMB	ZA
Гамбия	GMB	GA	Непал	NPL	NP	Зимбабве	ZWE	ZI

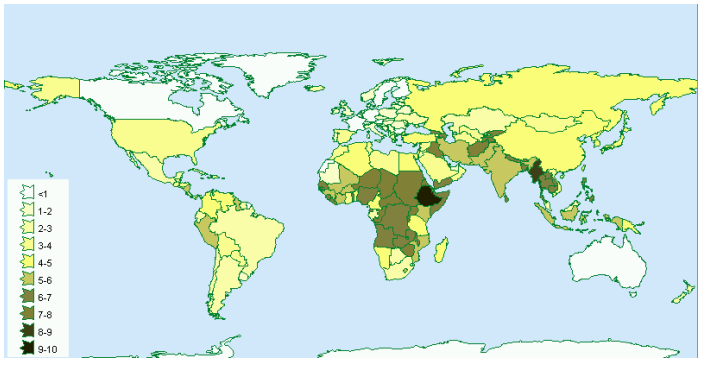
Рис. П2. Раскраска географической карты по пяти индексам



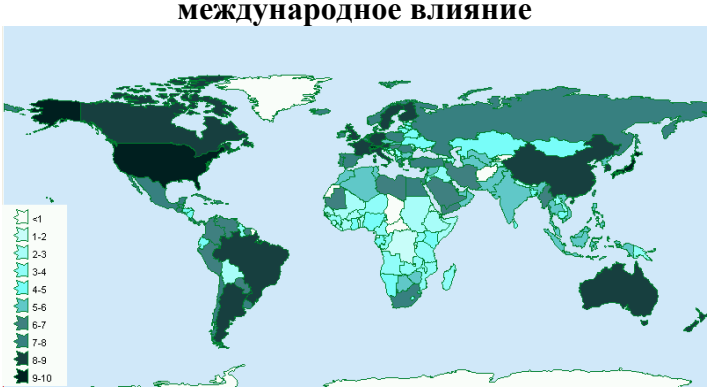
качество жизни



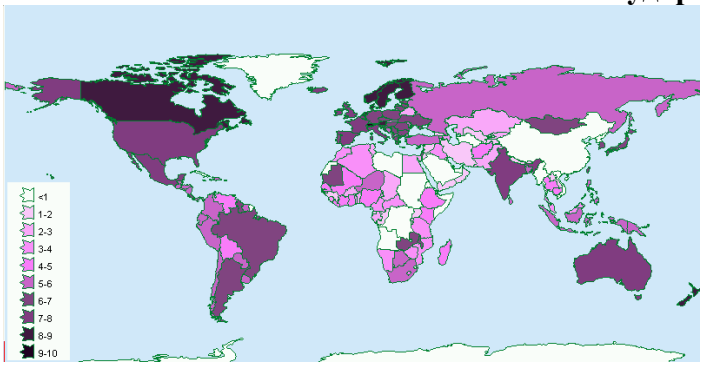
международное влияние



угрозы



Государственность



Демократичность

Рис. П3. Пример раскраски географической карты одновременно по трем индексам: демократичность в красном канале, международное влияние в зеленом и уровень жизни в синем

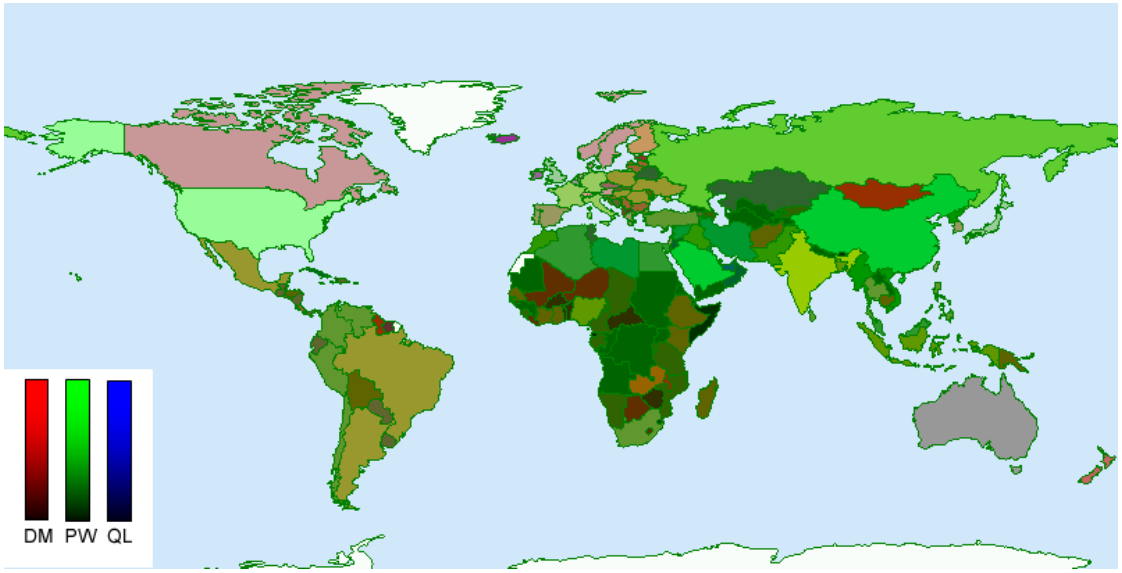
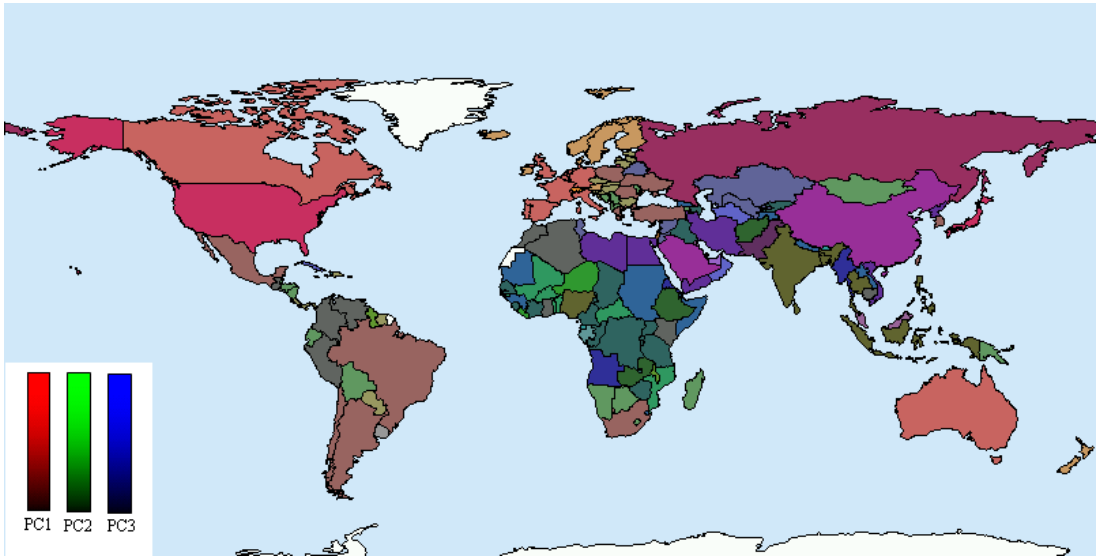
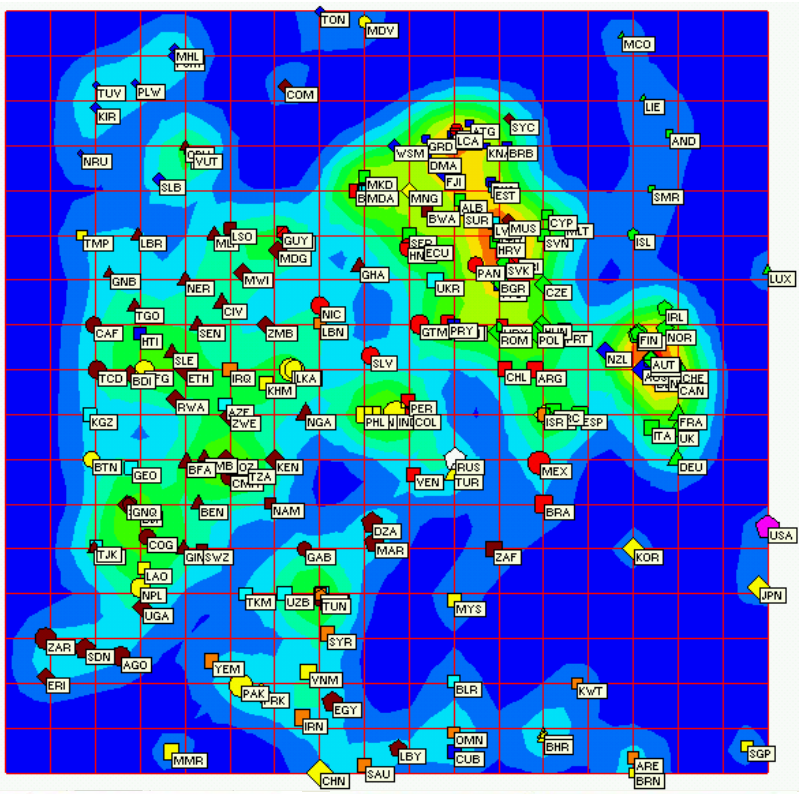
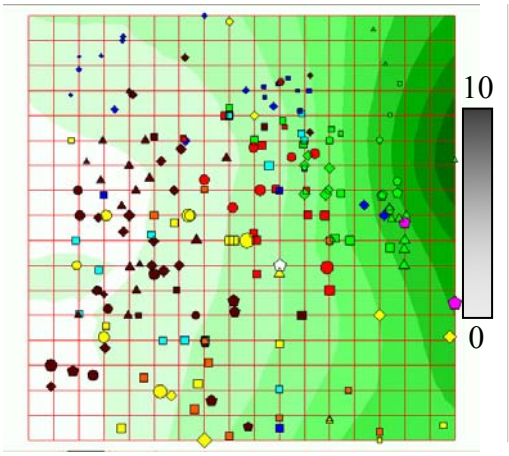


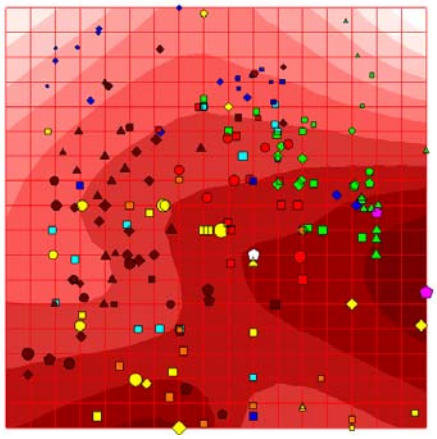
Рис. П4. Раскраска географической карты по трем первым главным компонентам первая компонента в красном канале, вторая в зеленом и третья в синем



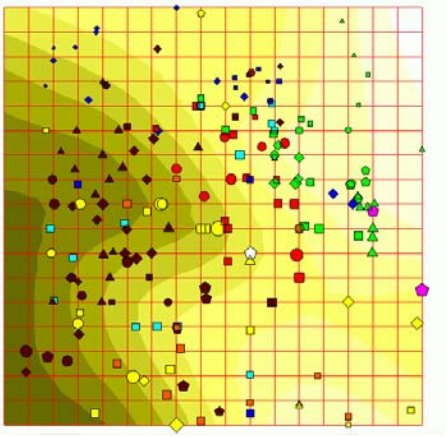




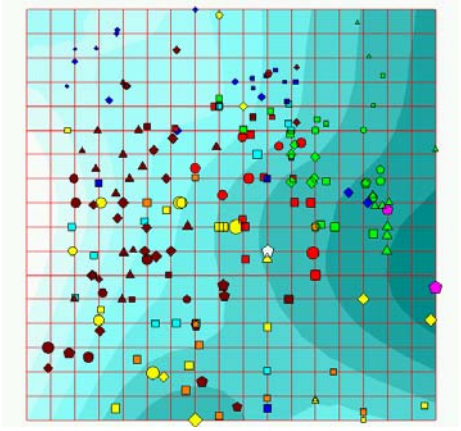
QL (качество жизни)



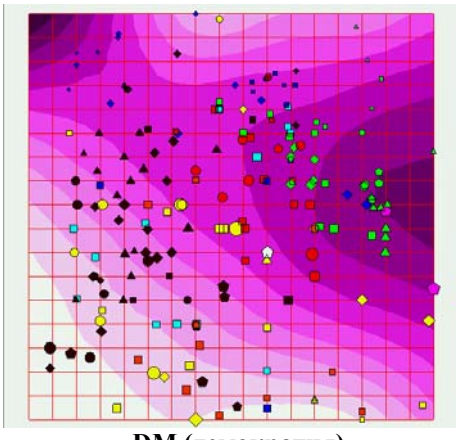
PW (влияние)



MN (угрозы)



SO (государственность)



DM (демократия)

	PC1	PC2	PC3	PC4	PC5
QL	0.55	-0.03	0.20	0.74	0.33
PW	0.31	-0.26	-0.69	-0.29	0.53
MN	-0.44	0.09	-0.64	0.59	-0.20
SO	0.51	-0.39	-0.19	-0.02	-0.75
DM	0.38	0.88	-0.22	-0.13	-0.14

Рис. П5. Карта данных индексов Политического Атласа Современности и атлас информационных раскрасок, построенные для «взвешенных данных». Каждой точке-стране приписан вес, равный *численности ее населения*. На раскрасках PW и SO явным образом проявляются демократический (во главе с США) и тоталитарный (во главе с Китаем) «полюса мировой силы». Внизу справа дана таблица весов новых главных компонент для взвешенных данных.