КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 4 С. 923–939

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 51-77

Эмпирическая проверка теории институциональных матриц методами интеллектуального анализа данных

И. Л. Кирилюк[1, a], А. И. Волынский[1], М. С. Круглова[1], А. В. Кузнецова[2,b], А. А. Рубинштейн[1], О. В. Сенько[3,c]

[1] Институт экономики РАН,
Россия, 117218, г. Москва, Нахимовский проспект, д. 32
[2] ИБХФ РАН, Россия, 119334, г. Москва, ул. Косыгина, д. 4
[3] ВЦ РАН, Россия, 119333, г. Москва, ул. Вавилова, д 40
E-mail: [a] igokir@rambler.ru, [b] azfor@narod.ru, [c] senkoov@mail.ru

Получено 20 февраля 2015 г.,
после доработки 8 апреля 2015 г.

Цель настоящего исследования состояла в установлении достоверной взаимосвязи показателей внешней среды и уровня освоенности территорий с характером доминирующих в странах институциональных матриц. Среди индикаторов внешних условий представлены как исходные статистические показатели, напрямую полученные из баз данных открытого доступа, так и сложные интегральные показатели, сформированные путем применения метода главных компонент. Оценка точности распознавания стран с доминированием X- или Y-институциональных матриц по перечисленным показателям проводилась с помощью ряда методов, основанных на машинном обучении. Была выявлена высокая информативность таких показателей, как освоенность территории, амплитуда осадков, летние и зимние температуры, уровень рисков.

Ключевые слова: теория институциональных матриц, машинное обучение

Empirical testing of institutional matrices theory by data mining

I. L. Kirilyuk[1], A. I. Volynsky[1], M. S. Kruglova[1], A. V. Kuznetsova[2], A. A. Rubinstein[1], O. V. Senko[3]

[1] Institute of Economics of RAS, 32 Nakhimovsky prospect, Moscow 117218 Russia
[2] Emanuel Institute of Biochemical Physics of RAS, 4 Kosygina str., Moscow, 119334, Russia
[3] Dorodnicyn Computing Centre of RAS, 40 Vavilov str., Moscow, 119333, Russia

Abstract. — The paper has a goal to identify a set of parameters of the environment and infrastructure with the most significant impact on institutional-matrices that dominate in different countries. Parameters of environmental conditions includes raw statistical indices, which were directly derived from the databases of open access, as well as complex integral indicators that were by method of principal components. Efficiency of discussed parameters in task of dominant institutional matrices type recognition (X or Y type) was evaluated by a number of methods based on machine learning. It was revealed that greatest informational content is associated with parameters characterizing risk of natural disasters, level of urbanization and the development of transport infrastructure, the monthly averages and seasonal variations of temperature and precipitation.

Keywords: institutional matrices theory, machine learning

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 4, pp. 923–939 (Russian).

Работа выполнена при поддержке гранта РГНФ (проект № 14-02-00422)

1. Введение

Разработанная С. Г. Кирдиной теория институциональных матриц утверждает, что характер развития общества определяется доминированием в его структуре одной из двух систем базовых институтов, или институциональной матрицы — либо X, либо Y [Кирдина, 2014]. X-матрица характеризуется институтами редистрибутивной экономики, унитарным политическим устройством, коммунитарной идеологией. Y-матрица характеризуется институтами рыночной экономики, федеративным политическим устройством, индивидуалистской идеологией. На рисунке 1 представлены теоретически предполагаемые страны X- и Y-матриц (серый и темно-серый цвет соответственно).

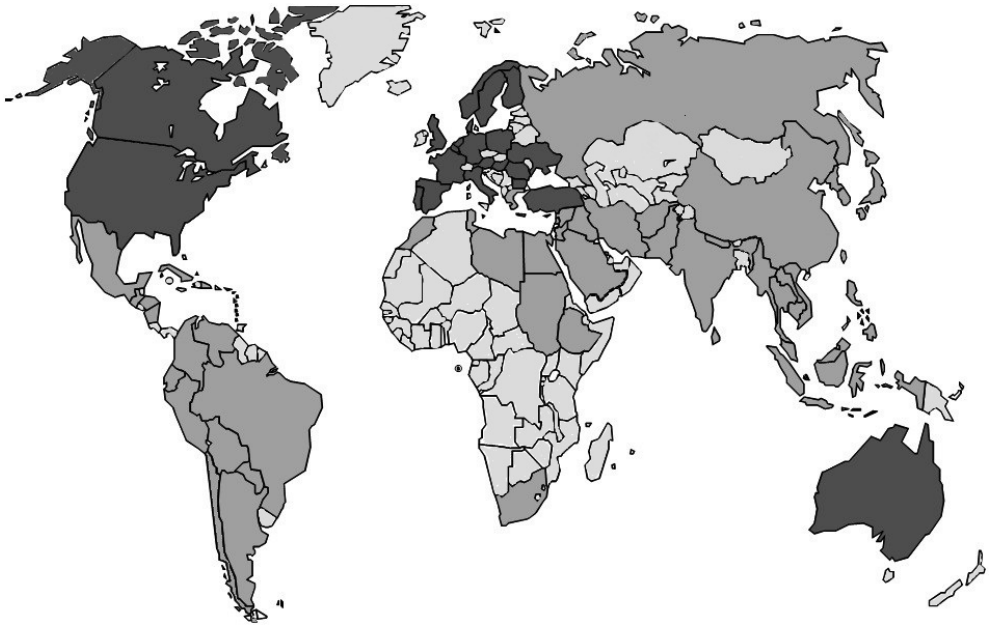

Рис. 1. Страны X-матрицы представлены серым цветом, страны Y-матрицы – темно-серым цветом. Светло-серым закрашены страны, не вошедшие в исследование

Несмотря на то, что положения этой теории активно используются в сравнительных исследованиях разных стран, а также при анализе особенностей социально-экономического развития России, эмпирической проверке собственно теории институциональных матриц уделяется несопоставимо меньшее внимание. Среди немногочисленных работ можно отметить исследование Н. В. Латовой [Латова, 2003], где автор проверяет выводы теории институциональных матриц с помощью этнометрической методологии Г. Хофстеде. В результате Латова подтверждает, что отнесенность стран к группе государств с доминированием той или иной институциональной матрицы совпадает с группировкой Хофстеде, в которой используются следующие культурные показатели: дистанция власти, индивидуализм, маскулинность [Hofstede, 1993].

Одной из задач проекта «Анализ макроинституциональной циклической динамики», реализуемого под руководством С. Г. Кирдиной (подробнее см. [Кирдина, 2015]), была эмпирическая проверка основополагающей гипотезы теории институциональных матриц о взаимосвязи внешних условий, в которых развивается государство, с характером доминирующей в нем матрицы. Среди показателей внешних условий основное внимание было уделено климатическим характеристикам. Также использовались показатели освоенности территории, природных рисков и др.

Методы многомерного статистического анализа (в том числе факторного, кластерного и дискриминантного анализа) применялись ранее для исследований социально-экономических систем стран мира другими исследователями, см., например, [Жуковская, Мучник, 1976]. Груп-

па ученых из МГИМО(У) МИД РФ провела и опубликовала ряд исследований по этой теме в рамках проекта «Политический атлас современности» [Мельвиль, 2006]. Особенность наших исследований в том, что мы применяем несколько иные современные подходы, подтвердившие свою эффективность и устойчивость на большом числе задач и различных областей знаний, для обоснования конкретной теории институциональных матриц. Использование метода оптимальных достоверных разбиений (ОДР) позволило не только наглядно выделить на диаграммах рассеяния области преобладания какого-либо из типов институциональных матриц, но также оценить достоверность различий между этими областями. Следует отметить, что метод ОДР основан на перестановочном тесте, не требующем предположений о типах распределений и не чувствительном к размерам выборок [Kuznetsova et al., 2014], что является актуальным для рассматриваемой задачи ввиду ограниченного числа сравниваемых эталонных стран. Наглядное выделение областей преобладания типов матриц в пространстве показателей облегчает анализ возможных причинно-следственных связей. Использование различных технологий распознавания позволило не только повысить точность прогноза, но также дать более подробную, основанную на различных критериях оценку сходства каждой из стран с эталонными группами, построить более устойчивое коллективное решение. Отметим, что повышение устойчивости является особенно важным в условиях малых выборок [Журавлев, Рязанов, Сенько, 2006].

2. Материалы и методы

2.1. Анализируемая информация

На первом этапе была сформирована выборка стран в количестве 70.

В выборку страны включались по следующим критериям. Они должны иметь достаточно продолжительную историю независимого существования и быть политически самостоятельными. Также критериями являются большие число жителей и размер территории.

Сначала был взят список из 50 стран, которые подписали устав ООН в 1945 году. Из этого списка затем были исключены: страны, которые потом распались на другие государства (Россия как преемница СССР была в базе оставлена), страны с населением менее 5 млн человек (Коста-Рика, Либерия, Люксембург, Уругвай), страны с площадью территории менее 30 тыс. кв. км (Сальвадор), а также Республика Гаити (из-за ее нестабильности). Затем в выборку были добавлены страны с населением более 5 млн чел., территорией не менее 30 тыс. кв. км и имеющие период независимости не менее 55 лет. Итоговый выборочный список включает 70 стран:

- **X-матрица**: Афганистан, Аргентина, **Белоруссия,** Болгария, Боливия, **Бразилия*,** Венгрия, **Венесуэла,** Вьетнам, Гватемала, Гондурас, Греция, Доминиканская Республика, **Египет**, Индия, Индонезия, Иордания, Ирак, Иран, Камбоджа, **КНДР, КНР**, Колумбия, **Куба**, **Лаос**, Ливан, Ливия, Малайзия, Марокко, **Мексика, Мьянма, Непал,** Никарагуа, Пакистан, Парагвай, **Перу,** Польша, Португалия, **Республика Корея**, **Российская Федерация,** Румыния, **Саудовская Аравия,** Сирия, Судан, Таиланд, Тунис, Турция, Украина, **Филиппины,** Чили, Шри-Ланка, Эквадор, Эфиопия, **Южно-Африканская Республика, Япония.**
- **Y-матрица**: Австралия, **Австрия, Бельгия, Великобритания, Дания, Германия, Испания, Италия**, Канада, **Нидерланды, Норвегия, США, Финляндия, Франция, Швеция.**

  * — жирным шрифтом выделены страны, вошедшие в эталонную выборку (31 страна: 18 — X-стран, 13 — Y) — для создания решающего правила при машинном обучении.

Предварительно имеющаяся информация свидетельствует о том, что стран, в которых доминируют X-институты, примерно в два раза больше, то есть 2/3 выборки.

Далее была сформирована база данных показателей внешних условий.

Предварительный список индикаторов, на основе которых можно судить о коммунальности/некоммунальности материально-технологической среды, включает в себя следующие группы:

1. *Уровень хозяйственных рисков.*
2. *Заселенность территории.*
3. *Транспортная инфраструктура.*
4. *Обеспеченность минеральными ресурсами.*
5. *Прочие характеристики.*

Статистические данные взяты на последнюю доступную дату для каждой страны, включенной в выборку.

Более 50 показателей взято из базы данных сайта www.worldbank.org.

Это данные о природных катаклизмах и ущербе от них, о числе погибших от катаклизмов, данные о температуре и осадках, водных ресурсах, урожайности зерновых и территории, засеянной зерновыми, грузообороте водного транспорта, населении городов, сельскохозяйственных и пахотных угодьях, о площади лесов.

С сайта www.cia.gov (база данных The World Factbook) взяты данные о населении, территории, годе обретения независимости, береговой линии, водных путях, обычных и железных дорогах, трубопроводах, запасах нефти и природного газа.

Также некоторые показатели взяты из источников www.gapminder.org/, http://faostat3.fao.org/, http://www.indexmundi.com, http://en.wikipedia.org, http://unstats.un.org, некоторые данные о запасах природных ресурсов взяты из http://www.world-nuclear.org/, http://www.bp.com/, http://minerals.usgs.gov/.

Часть показателей была также получена из исходных посредством арифметических преобразований, таких, например, как деление на площадь, или количество населения страны, или ВВП. Всего в расчетах участвовало 116 исходных и 15 показателей, полученных с помощью метода главных компонент. Все эти показатели прошли предварительный отбор на информативность методом оптимальных достоверных разбиений (ОДР) по индикатору двух групп эталонных стран. В конечную базу вошли только наиболее информативные показатели (см. таблицу 1).

Таблица 1. Краткое и полное наименование показателей, по которым проводился анализ, метод расчета и источники данных

Кратко	**Наименование показателя**	**Метод расчета, год сбора данных, способ получения показателя**
Max_t_град_ Цельсия	Максимальная температура за год, градусов Цельсия	Максимальная среднемесячная (из средних значений за 1961–1999 гг.) температура, градусов Цельсия
Min_t_град_ Цельсия	Минимальная температура за год, градусов Цельсия	Минимальная среднемесячная (из средних значений за 1961–1999 гг.) температура, градусов Цельсия
Амплитуда_ осадков	Среднегодовая амплитуда осадков, мм	Разница между максимальным и минимальным (за 12 месяцев) среднемесячными значениями выпавших осадков, мм, по значениям за 1961–1999 гг.
Амплитуда_ температур	Среднегодовая амплитуда температур, градусов Цельсия	Разница между максимальной и минимальной (из средних значений за 1961–1999 гг.) среднемесячной температурами, градусов Цельсия
Зимние_ осадки	Уровень осадков в холодные месяцы	Интегральный по 6 исходным показателям (среднемесячных значений осадков с ноября по март, мм по значениям за 1961–1999 гг.), полученный на основе метода главных компонент

Кратко	**Наименование показателя**	**Метод расчета, год сбора данных, способ получения показателя**
Зимние_ температуры	Температура в холодные месяцы	Интегральный по 7 исходным показателям (среднемесячных температур с октября по апрель 1961–1999 гг., а также средней по данным 12 месяцев температуры за год, градусов Цельсия), метод главных компонент
Городское_ население	Концентрация городского населения	Интегральный по 3 исходным показателям (доля населения крупнейшего города, %; доля населения столицы, %; доля населения в городах свыше 1 млн чел., %, 2012 г.), метод главных компонент
Летние_ осадки	Уровень осадков в теплые месяцы	Интегральный по 7 исходным показателям (среднемесячных значений осадков с мая по октябрь, а также значения осадков за год, мм, по значениям за 1961–1999 гг.), метод главных компонент
Летние_ температуры	Температуры в теплые месяцы	Интегральный по 5 исходным показателям (среднемесячных температур с мая по сентябрь 1961–1999 гг., градусов Цельсия), метод главных компонент
Ресурсы_ нефти_и_газа	Запасы углеводородов	Интегральный по 2 исходным показателям за 2013 г. (запасы нефти, млн баррелей в расчете на 1000 кв. км территории; запасы газа, миллиард кубометров в расчете на 1000 кв. км территории), метод главных компонент.
Ресурсы_ леса_и_воды	Обеспеченность водными и лесными ресурсами	Интегральный по 3 исходным показателям (доля территории, занятой лесами, %, 2011 г.; запасы пресной воды в кубометрах в расчете на 1000 кв. км; значение осадков за год, мм, 2011 г.), метод главных компонент
С/х_освоенность	Сельскохозяйственная освоенность территории	Интегральный по 3 исходным показателям (доля сельхозугодий, %, 2009–2011 гг.; доля пашни, %, 2011 г.; доля территории под зерновыми культурами, %, 2010–2012 гг.), метод главных компонент
Урбан+ транспорт	Освоенность территории	Интегральный по 3 исходным показателям (плотность автомобильных дорог, км на 1000 кв. км; плотность железных дорог; км на 1000 кв. км; доля городского населения, %, 2012), метод главных компонент. Данные о дорогах варьируют для разных стран в диапазоне 2000–2012 гг.
Уровень_ рисков	Уровень природных рисков	Интегральный по 3 исходным показателям за 1970–2008 гг., чел/100 000 чел. (доля населения, подверженного стихийным бедствиям; доля пострадавших от засухи; доля пострадавших от наводнений), метод главных компонент
Зерновые_ культуры	Урожайность зерновых культур	Показатель взят непосредственно из базы данных, размерность кг с гектара, 2012 г.

Большинство показателей таблицы 1 получено на основе базы данных Всемирного банка. Исключения: «Ресурсы_нефти_и_газа» рассчитаны на основе базы данных ЦРУ The World Factbook (https://www.cia.gov/library/publications/the-world-factbook/), там же взяты плотности автомобильных и железных дорог для показателя «Урбан + транспорт». «Уровень_рисков» рассчитан на основе базы данных Gapminder.org.

В первом столбце таблицы 1 приведены названия показателей, используемые в тексте, в среднем столбце — их полное название, а в правом столбце — история создания показателя и источник информации по нему со ссылкой.

Исследования связи типа доминирующей институциональной матрицы с перечисленными выше индикаторами проводились на выборке из 31 страны, для которых доминирующий тип

известен. Данная выборка (эталонная) далее будет называться также обучающей выборкой. Она будет использоваться для поиска закономерностей и настройки алгоритмов распознавания. В обучающую выборку включены 13 стран с Y-институциональными матрицами и 18 стран с X-институциональными матрицами — (далее X- и Y-страны).

2.2. Методы анализа

Исследования проводились с использованием методов интеллектуального анализа данных, включающих метод главных компонент, метод оптимальных достоверных разбиений, а также набор методов распознавания.

Метод главных компонент. Причина использования метода главных компонент в необходимости сокращения весьма обширного исходного набора показателей. Для групп однородных по своему смыслу показателей методом факторного анализа (метод выделения: анализ главных компонент; метод вращения: варимакс с нормализацией Кайзера) вычислялись соответствующие главные компоненты (с полной объясненной дисперсией более 90 процентов), что позволило перейти от анализа исходных таблиц, содержащих значения 116 индикаторов, к набору 11 интегральных показателей.

Метод оптимальных достоверных разбиений (ОДР).

Метод ОДР использовался для изучения связи типа институциональной матрицы с отдельными индикаторами или парами индикаторов. Метод ОДР основан на оптимальном разбиении интервалов значений независимых (объясняющих) переменных или совместных областей значений пар переменных [Senko, Kuznetsova, 2006]. Ищутся параллельные координатным осям границы, позволяющие наилучшим образом отделить объекты с различными уровнями значений прогнозируемой переменной. При этом разделяющая способность разбиения оценивается с помощью статистики критерия χ^2. Верификация закономерностей производится с помощью перестановочных тестов, позволяющих в случае двумерных моделей оценить значимость различий по каждой из двух используемых переменных [Kuznetsova et al., 2014]. Перестановочный тест основан на многократном повторении вычислений оптимальных разбиений на большом числе случайных таблиц, полученных из исходной таблицы с помощью случайных перестановок позиций индикаторов класса относительно фиксированных позиций объясняющих переменных.

Одним из преимуществ подхода является возможность выявления нелинейных эффектов. Преимуществом также является удобная наглядная форма представления результатов анализа в виде двумерных диаграмм.

Методы распознавания.

Методы распознавания использовались для многофакторного оценивания типа институциональных матриц. Алгоритм распознавания классифицирует произвольную страну в группу X или группу Y по большому набору индикаторов. На первом этапе индикаторы используются для вычисления оценок за группы. Оценкой объекта за группу (класс) в теории распознавания называется число, являющееся мерой сродства объекта к этой группе. Чем больше величина оценки объекта за какую-либо группу, тем увереннее этот объект может быть отнесен к ней. Таким образом, алгоритмы распознавания позволяют численно выразить меру сходства страны с группой X или с группой Y, ранжировать страны по отношению к группам. Окончательная классификация производится по величине оценок, рассчитанных на первом этапе. Например, объект может быть отнесен в группу, оценка за которую максимальна. Но могут быть объекты, занимающие промежуточное среднее положение между полюсами.

Важной характеристикой связи типа институциональных матриц с произвольным набором индикаторов является точность производимой по ним классификации, которая может быть эффективно оценена с помощью техники скользящего контроля. Эта техника дает объективную несмещенную оценку точности распознавания и заключается в последовательном удалении одного объекта из выборки, создания решающего правила по оставшимся объектам и классификации данного удаленного объекта. Так происходит со всеми объектами выборки. Доля пра-

вильно распознанных объектов в этой процедуре (отнесенных в свой класс) является оценкой точности распознавания.

При анализе сравниваемых классов были использованы методы, вошедшие в компьютерную систему «Распознавание» [Журавлев, Рязанов, Сенько, 2006]. Подробная информация о системе «Распознавание» доступна на сайте www.solutions-center.ru. Прогностическая способность использованных методов оценивалась с помощью метода скользящего контроля. Под прогностической способностью в данном случае понималось правильное отнесение страны к группе, в которую ее до этого поместил эксперт.

Распознавание проводили следующими методами, входящими в систему «Распознавание».

Линейный дискриминант Фишера. Классический статистический метод, основанный на поиске направления в многомерном пространстве признаков, вдоль которого достигается наилучшая разделимость распознаваемых классов.

Линейная машина. Построение линейного решающего правила, задаваемого с помощью линейных функций, которые строятся для каждого из классов. Вычисление коэффициентов линейной функции осуществляется через поиск максимальной совместной подсистемы неравенств, что соответствует максимизации числа правильно распознанных объектов [Журавлев, Рязанов, Сенько, 2006].

Многослойный перцептрон (нейронная сеть). Многослойная нейронная сеть, которая обучается с помощью метода обратного распространения ошибки. Метод многократно описан в литературе.

Логические закономерности. Метод основан на вычислении коллективных решений по системам логических закономерностей. Под логической закономерностью понимается область признакового пространства, задаваемого с помощью конъюнкции неравенств для отдельных признаков [Рязанов, 2007; Журавлев, Рязанов, Сенько, 2006].

Голосование по тупиковым тестам. Сравнение распознаваемого объекта с эталонными осуществляется по различным «информативным» подмножествам признаков. В качестве подобных подсистем признаков используются тупиковые тесты (или аналоги тупиковых тестов для вещественнозначных признаков) различных случайных подтаблиц исходной таблицы эталонов. Был предложен академиком РАН Ю.И. Журавлевым [Дмитриев, Журавлев, Кренделев, 1966].

Q ближайших соседей. Решение о классификации принимается по ближайшим соседям распознаваемого объекта. Сходство между объектами задается с помощью евклидовой метрики [Журавлев, Рязанов, Сенько, 2006].

Алгоритм вычисления оценок (АВО). Распознавание осуществляется на основе сравнения распознаваемого объекта с эталонными по различным наборам признаков и на основе использования процедур голосования. АВО был предложен академиком РАН Ю.И. Журавлевым в начале 70-х годов XX века [Журавлев, 1978]. В описании были отражены передовые для того времени и сохраняющие свою актуальность концепции решения задач распознавания. Решение о классификации объекта принимается с помощью анализа интегральных оценок близости объекта к эталонным объектам каждого из классов по системам опорных множеств — подмножеств полного набора признаков. Признаками принято называть всевозможные показатели, используемые для решения задачи распознавания. В нашем случае использовалась система опорных множеств, включившая всевозможные подмножества признаков из полного набора.

Метод опорных векторов. Изначально, с целью устранения неоднозначности, предлагалось производить разделение с помощью гиперплоскости, одинаково удаленной от двух параллельных гиперплоскостей, разделяющих классы. При этом выдвигалось требование максимизации «зазора» — расстояния между этими двумя гиперплоскостями. Далее были разработаны модификации метода, позволяющие строить оптимальные гиперплоскости с минимальным числом ошибочных классификаций в случаях отсутствия линейной разделимости. Наконец, была изобретена модификация, позволяющая строить также нелинейные разделяющие поверхности. Данная модификация основана на виртуальном переходе в новое гипотетическое пространство, в котором классы оказываются линейно разделимы. Подход основан на зависимости линейной гиперплоскости только от скалярных произведений векторов описаний объектов обу-

чающей выборки. Поэтому для осуществления перехода достаточно приравнять скалярные произведения векторов описания объектов в гипотетическом пространстве к ядерным функциям в исходном пространстве признаков [Chris, Burges, 1998].

Прототипом метода опорных векторов являлся метод «Обобщенный портрет», разработанный В. Н. Вапником и А. Я. Червоненкисом [Вапник, Червоненкис, 1974]. В современном варианте метод был предложен в работе [Cortes, Vapnik, 1995].

Метод статистически взвешенных синдромов — основан на принятии коллективных решений по системам синдромов двумерных областей признакового пространства, в котором преобладают объекты одного из распознаваемых классов. Области задаются с помощью границ, которые находятся через оптимальные разбиения интервалов значений признаков. Метод СВС впервые был предложен в работе «Распознавание нечетких систем по методу статистически взвешенных синдромов» [Кузнецов, Сенько, Кузнецова и др., 1996].

Метод мультимодельных статистически взвешенных синдромов.

Также был использован метод распознавания «Мультимодельные статистически взвешенные синдромы» [Senko, Kuznetsova, 2010]. Отличием этого метода от метода СВС является использование не только одномерных разбиений, но и одновременный поиск оптимальных разбиений по парам переменных. При этом используются модели, в которых разбиения на парах показателей задаются с помощью границ, параллельных координатным осям, а также прямых линий под произвольным углом и сдвигом, дающих наилучшее разделение наблюдений из исследуемых классов.

3. Результаты

Метод **оптимальных достоверных разбиений** (ОДР) помог найти границы показателей, разделяющие интервал значений таким образом, чтобы с каждой стороны от границы были преимущественно наблюдения одного из исследуемых классов. Для каждого показателя оценивали эффективность найденной границы по специальному функционалу и определяли значимость каждого разбиения по соответствующему p-значению.

Таблица 2. Наиболее информативные показатели при одномерных разбиениях. Расположены по убыванию значимости

Показатель	Граница	Ниже границы		Выше границы		Функционал	p-значение
		Y	X	Y	X		
Урбан+транспорт	–0,1776	0	13	13	5	15,648146	0,0005
Амплитуда_осадков	85,28	13	5	0	13	15,648146	0,001
Летние_температуры	–0,3966	12	5	1	13	12,282372	0,0025
Зимние_температуры	0,08505	13	7	0	11	11,916664	0,0040
Min_t_градусов_Цельсия	10,05	13	7	0	11	11,916664	0,001
Уровень_рисков	–0,5710	11	4	2	14	11,386751	0,0035
Max_t_градусов_Цельсия	21,7999	13	8	0	10	10,317460	0,011
Летние_осадки_	0,3111	13	9	0	9	8,863636	0,0185
Урожайность_зерновых	5058,5	3	14	10	4	8,825682	0,0245
Зимние_осадки	–0,680	0	8	13	10	7,536231	0,0465
Амплитуда_температур	10,812950	0	8	13	10	7,536231	0,047

Сначала применяли метод ОДР для **одномерных показателей**. В таблице 2 приведены названия показателей по убыванию информативности с точки зрения разделения стран X и стран Y.

Далее приведены границы разбиения по каждому показателю. В последующих столбцах показано число наблюдений X-стран, находящихся ниже границы, число наблюдения Y-стран — ниже границы, и аналогично число X и Y выше границы. Два последних столбца показывают функционал, по которому делают вывод о значимости показателя и достоверном p-значении.

Пример одномерного разбиения приведен на рисунке 2.

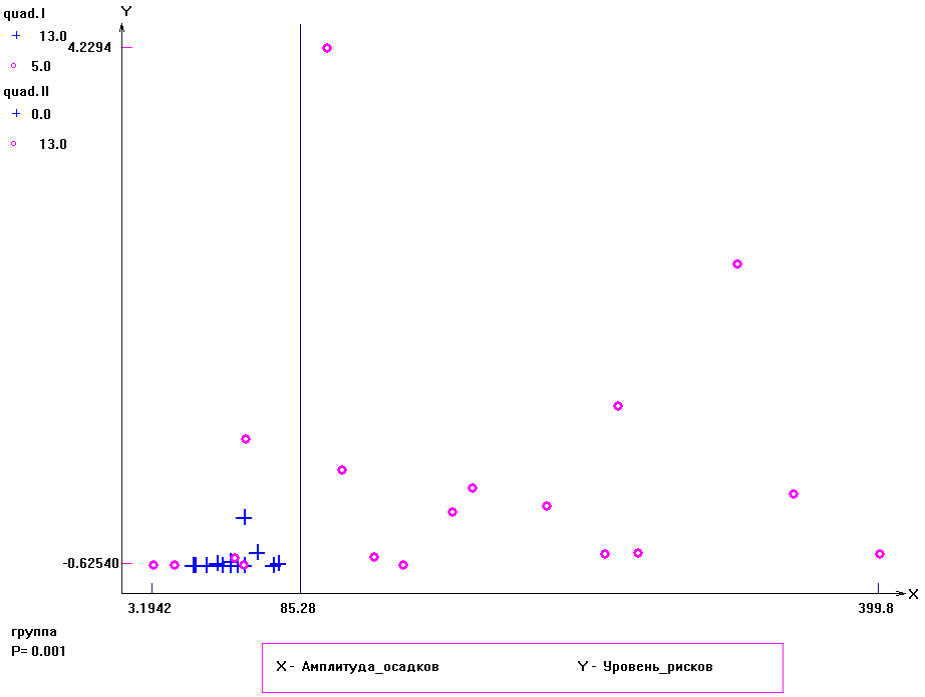

Рис. 2. По оси X — «Амплитуда осадков». По оси Y — «Уровень рисков», этот показатель взят только для развертки и роли не играет. Красные кружки означают наблюдения стран X. Синие крестики — наблюдения стран Y

На диаграмме рассеяния видно, что высокие значения амплитуды осадков (выше 85,28) соответствуют только странам из группы X (13 наблюдений). В этой области нет ни одного наблюдения из группы стран Y. И, наоборот, для всех стран из группы Y (13 наблюдений) амплитуды осадков имеют значения ниже границы. Хотя в этой области присутствует и 5 наблюдений из группы X. Достоверность разбиения, посчитанная с помощью перестановочного теста с использованием 3000 случайных перестановок индикатора класса, на уровне $p = 0{,}001$.

Двухмерный вариант метода ОДР.

Далее метод ОДР был применен на парах показателей. Аналогично одномерным разбиениям для каждого показателя была найдена граница разбиений, чтобы с обеих сторон от границы преобладали объекты или стран X или стран Y. Но в данном случае рассматривали уже квадранты, получающиеся при пересечении двух границ (см. таблицу 3).

На диаграмме рассеяния (рис. 3) хорошо выражены области, в которых преимущественно находятся наблюдения какого-то одного из исследуемых классов. Так, закономерности для стран Y (синие крестики) отражены в II квадранте: высокая урбанизация и развитый транспорт (выше –0,1776) при низком объеме амплитуды осадков (ниже 85,28). Достоверная значимость разбиения на уровне $p_x = 0{,}0145$. В то время как для X-стран, наоборот, характерны низкие значения интегрального показателя «Освоенность территории (Урбанизация + транспорт)» и высокие значения показателя «Амплитуды осадков» — квадранты I, II и IV. Достоверная значимость разбиения на уровне $p_y = 0{,}014$.

Из 15 показателей, вошедших в анализируемую базу, на одномерных разбиениях информативными оказались 11 показателей с достоверной значимостью $p < 0{,}05$.

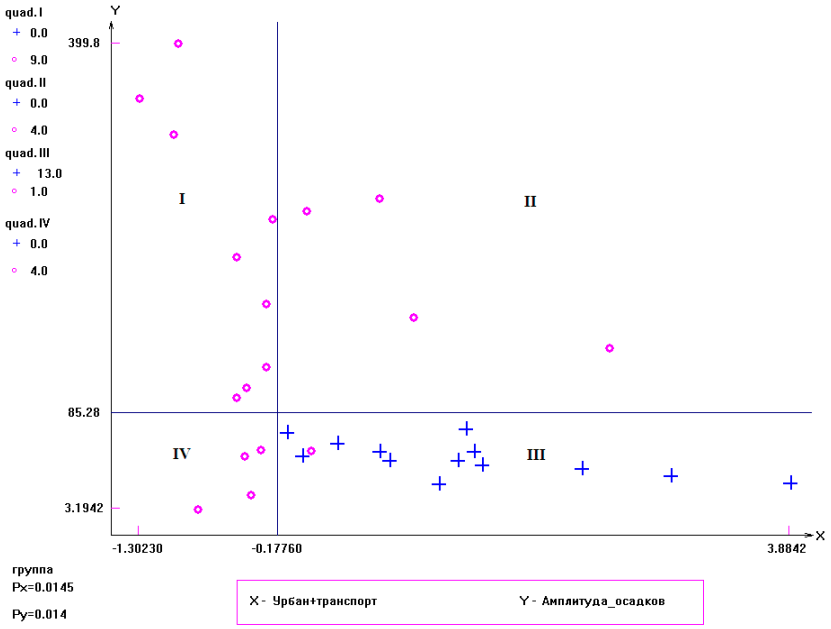

Рис. 3. По оси X значения интегрального показателя «Освоенность территории (урбанизация и транспорт)». По оси Y — исходный показатель «Амплитуда осадков». Красные кружки означают наблюдения стран X. Синие крестики — наблюдения стран Y

Таблица 3. Наиболее информативные показатели из достоверно значимых двумерных разбиений (за исключением показателей, вошедших в одномерные разбиения)

1-й показатель	**2-й показатель**	**Граница 1**	**Граница 2**	**px**	**py**
Ресурсы_нефти_и_газа	Амплитуда_осадков	–0,2452	53,6	0,001	0,038
Уровень_рисков	Сельскохозяйственная_освоенность	–0,1428	–0,1373	0,0685	0,025
Ресурсы_леса_и_воды	Летние_температуры	–0,1736	–1,1073	0,018	0,0215
Сельскохозяйственная_освоенность	Min_t_градусов_Цельсия	0,2334	3,3	0,035	0,01
Ресурсы_леса_и_воды	Сельскохозяйственная_освоенность	–0,1736	0,2334	0,037	0,0395
Ресурсы_леса_и_воды	Урожайность_зерновых	0,77325	3757,0	0,0195	0,0185

При двумерных разбиениях — на парах признаков — добавились такие показатели, как «Ресурсы нефти и газа», «Сельскохозяйственная освоенность», «Ресурсы леса и воды». Показатель «Концентрация городского населения» в оба набора не вошел.

На диаграммах рассеяния представлены наиболее информативные с точки зрения разделения стран на категории X и Y пары показателей. По каждому из показателей автоматически поставлены границы разбиения таким образом, чтобы с обеих сторон оставались наблюдения преимущественно одного класса.

На рисунке 4 можно видеть, что в верхних и правых квадрантах (I, II и III) находятся исключительно наблюдения стран X (красные кружки). Это означает, что для данных стран интегральный показатель уровня рисков выше значения (–0,1428). Достоверная значимость по данному показателю оценивается на уровне $p_x = 0{,}035$. Летние температуры выше интегрального значения (–0,089), $p_y = 0{,}017$. Для стран Y все наблюдения ограничены областью нижнего, левого квадранта (IV). То есть их значения ниже обеих границ. Следовательно, для этих стран характерны низкий уровень рисков и низкие летние температуры.

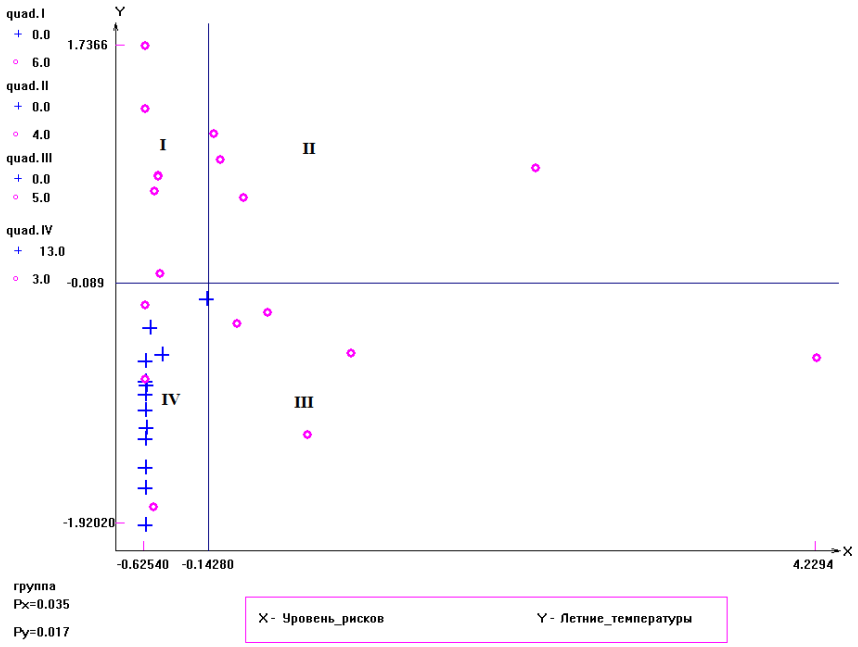

Рис. 4. По оси X — значения интегрального показателя «Уровень рисков». По оси Y — летние температуры. Красные кружки означают наблюдения стран X. Синие крестики — наблюдения стран Y

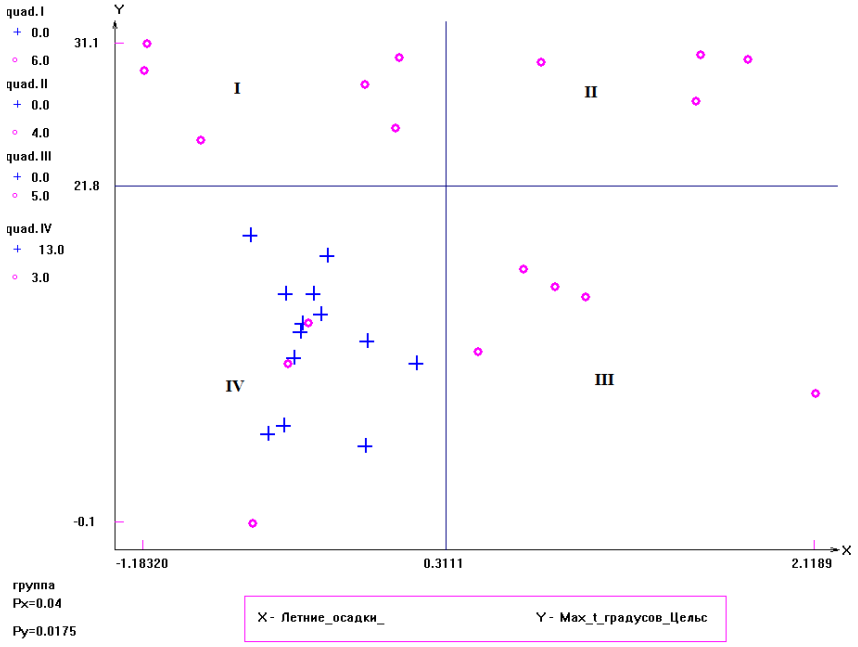

Рис. 5. Диаграмма рассеяния для показателей «Летние осадки» и «Максимальная температура по Цельсию». Обозначения, как на рисунке 2

На данной диаграмме рассеяния (рис. 5) группа стран Y опять занимает нижнее левое положение. То есть для них характерны значения производного интегрального показателя «Летние осадки» меньше 0,31 и значения исходного показателя «Максимальные температуры по Цельсию» ниже 21,8. Страны X занимают преимущественно верхние правые квадранты — выше границ. Три наблюдения класса X, попавшие в IV квадрант, вероятно, и дают ошибки при распознавании.

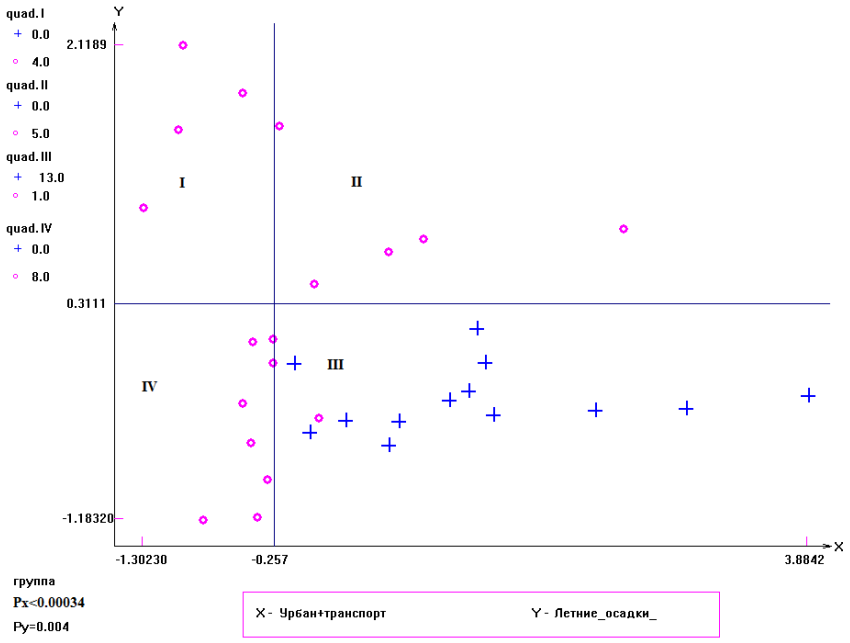

Рис. 6. Освоенность территории (урбанизация и транспорт) по оси X, летние осадки по оси Y. Обозначения, как на рисунке 2

На диаграмме рассеяния (рис. 6) отражена найденная методом ОДР закономерность для стран Y: высокая урбанизация и развитый транспорт (интегральный показатель выше –0,257) при низких значениях производного показателя «Летние осадки» (ниже 0,31). Для X-стран, наоборот, характерны низкие значения урбанизации и высокие объемы летних осадков.

Методы распознавания были применены для анализа эталонной (обучающей) выборки, содержащей 15 показателей (интегральных и исходных) по 31 стране. Наиболее эффективными оказались следующие методы распознавания: алгоритм вычисления оценок (АВО), метод опорных векторов (МОВ, support vector machine, SVM), метод статистически взвешенных синдромов, основанный на методе оптимальных разбиений (СВС), а также его модификация — мультимодельный метод статистически взвешенных синдромов (МСВС). Результаты распознавания на скользящем контроле представлены в таблицах 4 и 5.

Таблица 4. Эффективность распознавания методов интеллектуального анализа данных из системы «Распознавание» и авторских методов распознавания

№	Название метода	Правильно из 31	Точность распознавания, %
1	**Алгоритмы вычисления оценок (АВО)***	**28**	**90,3**
2	**Метод опорных векторов (SVM)**	**27**	**87,1**
3	**Статистически взвешенные синдромы (СВС)**	**26**	**83,9**
4	Многослойный перцептрон (нейронная сеть)	25	80,6
5	Мультимодельный метод статистически взвешенных синдромов (МСВС)	25	80,6
6	Линейная машина	25	80,6
7	Голосование по тупиковым тестам	24	77,4
8	Линейный дискриминант Фишера	24	77,4
9	Логические закономерности	22	71,0
10	Q ближайших соседей	20	64,5

* Жирным шрифтом выделены наиболее успешные методы.

Наилучший результат распознавания показал метод «Алгоритмы вычисления оценок». При этом правильно было распознано в классе Y — 12 из 13 стран (92,3 %), в классе X — 16 из 18 стран (88,9 %), ошибочно отнесено три страны (Финляндия, Белоруссия и КНР). Метод опорных векторов и метод статистически взвешенных синдромов дали только 4 и 5 ошибок из 31 страны соответственно.

Таблица 5. Результаты скользящего контроля на обучающей выборке. Использованные методы: алгоритмы вычисления оценок (АВО), метод статистически взвешенных синдромов (СВС), его мультимодельная модификация (МСВС), метод опорных векторов (МОВ)

Страна/ метод	Класс	АВО		СВС		МСВС		МОВ	
Австрия	**Y**	–0,011	**Y**	0,488	**Y**	0,323	**Y**	0,391	**Y**
Бельгия	**Y**	–0,065	**Y**	0,112	**Y**	0,165	**Y**	0,235	**Y**
Дания	**Y**	–0,066	**Y**	0,112	**Y**	0,146	**Y**	0,269	**Y**
Финляндия	**Y**	**0,047**	**X**	0,367	**Y**	**0,543**	**X**	0,482	**Y**
Франция	**Y**	–0,128	**Y**	0,218	**Y**	0,204	**Y**	0,321	**Y**
Германия	**Y**	–0,114	**Y**	0,112	**Y**	0,145	**Y**	0,257	**Y**
Италия	**Y**	–0,092	**Y**	**0,860**	**X**	**0,523**	**X**	0,407	**Y**
Нидерланды	**Y**	–0,209	**Y**	0,112	**Y**	0,165	**Y**	0,247	**Y**
Норвегия	**Y**	–0,079	**Y**	0,367	**Y**	0,420	**Y**	0,472	**Y**
Испания	**Y**	–0,061	**Y**	**0,960**	**X**	**0,621**	**X**	**0,538**	**X**
Швеция	**Y**	–0,048	**Y**	0,167	**Y**	0,279	**Y**	0,410	**Y**
Великобритания	**Y**	–0,141	**Y**	0,112	**Y**	0,269	**Y**	0,308	**Y**
США	**Y**	–0,049	**Y**	**0,836**	**X**	0,475	**Y**	0,433	**Y**
Белоруссия	**X**	**–0,018**	**Y**	**0,064**	**Y**	**0,155**	**Y**	**0,385**	**Y**
Бразилия	**X**	0,839	**X**	0,999	**X**	0,945	**X**	0,684	**X**
КНР	**X**	**–0,041**	**Y**	0,847	**X**	**0,478**	**Y**	0,581	**X**
Куба	**X**	0,346	**X**	0,993	**X**	0,889	**X**	0,804	**X**
КНДР	**X**	1,030	**X**	0,763	**X**	0,684	**X**	0,644	**X**
Египет	**X**	0,456	**X**	0,918	**X**	0,719	**X**	0,668	**X**
Япония	**X**	0,006	**X**	0,755	**X**	**0,461**	**Y**	**0,385**	**Y**
Лаос	**X**	0,169	**X**	0,999	**X**	0,968	**X**	0,906	**X**
Мексика	**X**	12,480	**X**	0,999	**X**	0,921	**X**	0,716	**X**
Мьянма	**X**	0,226	**X**	0,997	**X**	0,964	**X**	0,800	**X**
Непал	**X**	0,161	**X**	0,944	**X**	0,775	**X**	0,726	**X**
Перу	**X**	0,353	**X**	0,999	**X**	0,944	**X**	0,640	**X**
Филиппины	**X**	0,487	**X**	0,999	**X**	0,931	**X**	0,804	**X**
Республика Корея	**X**	0,030	**X**	0,600	**X**	0,550	**X**	0,548	**X**
Российская Федерация	**X**	0,009	**X**	0,696	**X**	0,531	**X**	**0,258**	**Y**
Саудовская Аравия	**X**	0,242	**X**	0,972	X	0,760	**X**	0,770	**X**
ЮАР	**X**	0,009	**X**	0,943	X	0,553	**X**	0,552	**X**
Венесуэла	**X**	0,070	**X**	0,999	X	0,968	**X**	0,793	**X**

* Жирным шрифтом выделены значения ошибочного распознавания.

Аналогично теми же методами был проведен прогноз типа институциональной матрицы для 39 стран, для которых не было экспертно установленного класса. При прогнозировании использовали только методы, которые дали наилучший результат распознавания на обучающей выборке: АВО, СВС, МСВС и МОВ.

Таблица 6. Результаты распознавания стран, не вошедших в обучающую выборку

Страна/ метод	**АВО**		**СВС**		**МСВС**		**МОВ**	
Афганистан	0,088	X	0,865	X	0,522	X	0,744	X
Аргентина	0,084	X	0,657	X	0,409	**Y***	0,481	**Y**
Австралия	0,386	X	0,993	X	0,745	X	0,708	X
Боливия	0,479	X	0,999	X	0,918	X	0,742	X
Болгария	–0,002	**Y**	0,276	**Y**	0,278	**Y**	0,427	**Y**
Камбоджа	0,488	X	0,999	X	0,963	X	0,942	X
Канада	0,331	X	0,586	X	0,429	**Y**	0,350	**Y**
Чили	–0,059	**Y**	0,211	**Y**	0,308	**Y**	0,397	**Y**
Колумбия	0,489	X	0,999	X	0,966	X	0,844	X
Доминиканская Республика	0,473	X	0,999	X	0,962	X	0,792	X
Эквадор	0,480	X	0,999	X	0,951	X	0,707	X
Эфиопия	0,462	X	0,999	X	0,907	X	0,833	X
Греция	–0,013	**Y**	0,937	X	0,578	X	0,551	X
Гватемала	0,478	X	0,999	X	0,962	X	0,889	X
Гондурас	0,489	X	0,999	X	0,964	X	0,856	X
Венгрия	–0,085	**Y**	0,331	**Y**	0,242	**Y**	0,405	**Y**
Индия	0,471	X	0,996	X	0,876	X	0,873	X
Индонезия	0,477	X	0,999	X	0,950	X	0,738	X
Иран	0,315	X	0,987	X	0,748	X	0,758	X
Ирак	0,441	X	0,995	X	0,833	X	0,809	X
Иордания	0,354	X	0,997	X	0,854	X	0,785	X
Ливан	0,003	X	0,978	X	0,717	X	0,700	X
Ливия	0,454	X	0,995	X	0,841	X	0,808	X
Малайзия	0,481	X	0,999	X	0,958	X	0,760	X
Марокко	0,044	X	0,987	X	0,716	X	0,640	X
Никарагуа	0,478	X	0,999	X	0,962	X	0,892	X
Пакистан	0,316	X	0,997	X	0,858	X	0,729	X
Парагвай	0,396	X	0,999	X	0,918	X	0,745	X
Польша	–0,097	**Y**	0,121	**Y**	0,167	**Y**	0,312	**Y**
Португалия	–0,045	**Y**	0,937	X	0,599	X	0,531	X
Румыния	–0,095	**Y**	0,211	**Y**	0,236	**Y**	0,444	**Y**
Шри Ланка	0,483	X	0,996	X	0,941	X	0,837	X
Судан	0,462	X	0,997	X	0,879	X	0,897	X
Сирия	0,173	X	0,987	X	0,678	X	0,732	X
Таиланд	0,486	X	0,999	X	0,954	X	0,911	X
Тунис	0,111	X	0,997	X	0,847	X	0,673	X
Турция	–0,075	**Y**	0,331	**Y**	0,305	**Y**	0,477	**Y**
Украина	–0,056	**Y**	0,211	**Y**	0,298	**Y**	0,362	**Y**
Вьетнам	0,481	X	0,999	X	0,928	X	0,842	X

* Жирным шрифтом выделены отнесения каким-то из методов к классу Y.

Режимы, примененные в данном анализе:

МСВС: порог при отборе закономерностей на величину статистики критерия χ^2 — porogXUU = 5,0; porogCOMPL = 0,2 — коэффициент регулирует отбор разбиений при возрастании сложности модели; Recognition Cross Validation — режим использования скользящего контроля.

Метод опорных векторов (SVM): Тип потенциальной функции — Гауссиана, параметр потенциальной функции — 11,0.

4. Обсуждение

Почти все показатели, вошедшие в базу данных, оказались высоко информативными. Это свидетельствует о высокой эффективности применения метода главных компонент. Исключение составил показатель «Концентрация городского населения».

Наиболее информативными показателями оказались: «Освоенность территории» (Урбан + транспорт) — $p < 0{,}0005$, «Амплитуда_осадков» — $p < 0{,}001$, «Летние_температуры» — $p < 0{,}0025$, «Зимние_температуры» — $p < 0{,}004$, «Min_t_градусов_Цельсия» — $p < 0{,}001$, «Уровень_рисков» — $p < 0{,}0035$.

Надо отметить характер расположения Y-стран на диаграммах рассеяния (см. рис. 2–6) — они располагаются значительно компактнее, с меньшим разбросом наблюдений, чем у стран X. То есть Y-способ организации обусловлен более благоприятными внешними условиями.

При этом для стран Y отмечены значения по освоенности территории выше границы разбиения (> –0,1776), а для стран X — ниже (см. таблицу 2). Что касается остальных климатических показателей, то страны Y занимают преимущественно положения с умеренными значениями (ниже границы) по амплитуде осадков (< 85,28 мм), по летним (< –0,3966) и зимним температурам (< 0,085), по уровню рисков (< –0,571), по минимальным (< 10,05° С) и максимальным (< 21,8° С) температурам, а также по летним осадкам (< 0,3). Выше границы страны Y по урожайности зерновых (> 5058 кг/га), по зимним осадкам (> –0,680) и амплитуде температур (> 10,8).

Для большинства стран, не вошедших в группу эталонных и не используемых при обучении, все четыре метода распознавания дают одинаковый результат, что говорит о высокой точности отнесения к определенному классу. Но есть ряд стран, которые разные методы относят к разным классам. Такие страны, как Греция и Португалия, методом АВО отнесены в класс Y, а остальными тремя методами — в класс X. Аргентина и Канада отнесены в класс X методами АВО и СВС, а двумя другими методами в класс Y. Можно сделать заключение, что только климатических показателей и информации об освоенности территории недостаточно для того, чтобы однозначно относить каждую страну к тому или иному классу базовых институциональных матриц.

Выводы

Результаты данной работы объективно подтверждают наличие статистически значимой связи между спецификой внешних условий существования государств и характером доминирующих в них институциональных матриц. Наиболее информативными показателями с точки зрения взаимосвязи климата и освоенности территории с особенностями матриц базовых институтов стали «Освоенность территории» (урбанизация + транспорт), амплитуда осадков, летние температуры, зимние температуры, минимальная температура (°С), уровень рисков.

Для стран с Y-институциональной матрицей характерно компактное расположение наблюдений, в то время как разброс значений климатических показателей для X-страны намного шире.

Использованные в работе методы теории распознавания позволяют работать со сложными социально-экономическими объектами неоднозначной природы. Определены методы, позво-

ляющие распознавать государства с определенными институциональными матрицами (X или Y) наилучшим образом: это алгоритмы вычисления оценок (ABO), метод опорных векторов (MOB), метод статистически взвешенных синдромов (СВС).

Перспективы работы связаны с выявлением дополнительных показателей, которые, вероятно, смогут сделать процедуру распознавания более эффективной.

Благодарности

Авторы выражают искреннюю благодарность д. соц. н., к. э. н. С. Г. Кирдиной, академику РАН, д. э. н. В. И. Маевскому, д. т. н. С. Ю. Малкову, д. х. н. Ю. Л. Словохотову за полезные обсуждения.

Список литературы

Вапник В. Н., Червоненкис А. Я. Теория распознавания образов (статистические проблемы обучения). — М. : Наука, 1974. — 416 с.

Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. О математических принципах классификации предметов и явлений // Дискретный анализ. — 1966. — Вып. 7. — Новосибирск: ИМ СО АН СССР. — С. 3–11.

Жуковская В. М., Мучник И. Б. Факторный анализ в социально-экономических исследованиях. — М. : Статистика, 1976.

Журавлев Ю. И., Никифоров В. В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — С. 1–11.

Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Вып. 33. — С. 5–68.

Журавлев Ю. И. Избранные научные труды. — М. : Магистр, 1998. — 420 с.

Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М. : Фазис, 2006. — 159 с.

Кирдина С. Г. Институциональные матрицы и развитие России. Введение в X-Y-теорию. Изд. 3-е, перераб., расш. и иллюстр. — М.–СПб. : Нестор-История, 2014. — 468 с.

Кирдина С. Г. К анализу макроинституциональной циклической динамики. В: Эволюция экономической теории: воспроизводство, технологии, институты // Материалы X международного симпозиума по эволюционной экономике, г. Пущино 12–14 сентября 2013 г. / Отв. ред. В. И. Маевский и С. Г. Кирдина. СПб. : Алетейя, 2015.

Кузнецов В. А., Сенько О. В., Кузнецова А. В. и др. Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической характеристики нормы и хронической патологии // Химическая физика. — 1996. — Т. 15, № 1. — С. 81–100.

Латова Н. В. В какой матрице мы живем? (Этнометрическая проверка теории институциональных матриц) / Экономический вестник Ростовского государственного университета. — 2003. — Т. I, № 3. — С. 89–94.

Мельвиль А. Ю. «Политический атлас современности»: замысел и общие теоретико-методологические контуры проекта // Полис. — 2006. — № 5. — С. 6–14.

Рязанов В. В. Логические закономерности в задачах распознавания (параметрический подход) // Ж. вычисл. матем. и матем. физ. — 2007. — Т. 47, № 10. — С. 1793–1808.

Chris. J. C. Burges A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands // Appeared in: Data Mining and Knowledge Discovery. — 1998. — 2. — P. 121–167.

Cortes C. Vapnik V. Support-vector networks // Machine Learning. — 1995. — Vol. 20 (3): 273.

Hofstede G. (March 1993). Cultures and Organizations: Software of the Mind. Administrative Science Quarterly (Johnson Graduate School of Management, Cornell University) 38 (1): 132–134.

Kuznetsova A. V., Kostomarova I. V., Sen'ko O. V. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients // Pattern Recognition and Image Analysis. — 2014. — Vol. 24, Is. 1. — P. 114–123.

Senko O. V., Kuznetsova A. V. The Optimal Valid Partitioning Procedures // «InterStat», 39 of pages, http://interstat.statjournals.net/YEAR/2006/articles/0604002.pdf .

Senko O. B., Kuznetsova A. B. A recognition method based on collective decision making using systems of regularities of various types // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 2. — P. 152–162.